HELSINGIN YLIOPISTO

Lexical coverage in ELF

Christine Stevenage
Master's thesis
English Philology
Department of Languages
University of Helsinki
February 2018

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty **Faculty of Arts**	Laitos – Institution – Department **Department of Languages**	
Tekijä – Författare – Author **Christine Stevenage**		
Työn nimi – Arbetets titel – Title **Lexical coverage in ELF**		
Oppiaine – Läroämne – Subject **English Philology**		
Työn laji – Arbetets art – Level **Master's Thesis**	Aika – Datum – Month and year **February 2018**	Sivumäärä– Sidoantal – Number of pages **64 pages**

Tiivistelmä – Referat – Abstract

The aim of this study was to determine how much vocabulary is needed to understand English in contexts where it is spoken internationally as a lingua franca (ELF). This information is critical to inform vocabulary size targets for second language (L2) learners of English. The current research consensus, based on native-English-speaker data, is that 6,000–7,000 word families plus proper nouns are needed. However, since English has become a global lingua franca, native speakers of English have become a minority: in fact, today, there are around two billion speakers of English worldwide, of which less than a quarter are native speakers. This means that non-native speakers of English are more likely to interact with other non-native speakers than with native speakers. Thus, using findings based on solely native-speaker data may not provide the most accurate information needed to inform vocabulary size targets for L2 learners of English. Indeed, this information needs to be supplemented with data from competent non-native speakers of English who can represent a legitimate model for L2 learners of English.

This study uses the largest freely available corpus of general, spoken ELF in Europe: the one million-word Vienna-Oxford International Corpus of English (VOICE). The word family was used as a lexical counting unit, and the lexical coverage of VOICE was calculated for various thresholds of the most frequent word families in the corpus. A comparative analysis was carried out to determine the lexical coverage of VOICE provided by frequency ranked word lists based on data from the British National Corpus of English and the Contemporary Corpus of American English.

The main findings of this study indicate that fewer than 3,000–4,000 word families plus proper nouns can provide the lexical resources needed to understand English in international contexts where it is spoken as a lingua franca. This is approximately half the number of word families (i.e. 6,000–7,000 word families plus proper nouns) which scholars have claimed are needed to understand spoken English. The findings of this study represent a substantial saving in vocabulary size targets for L2 learners of English who wish to be functional in understanding English spoken as an international lingua franca.

Avainsanat – Nyckelord – Keywords
English as a Lingua Franca (ELF), Vienna-Oxford International Corpus of English (VOICE), Vocabulary, Lexical Coverage, Learning English as a Foreign Language (EFL)

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgfter – Additional information:
Full list of word families form VOICE available from the author upon request.

Contents

Abbreviations

BNC: British National Corpus

CANCODE: Cambridge and Nottingham Corpus of Discourse in English

COCA: Corpus of Contemporary American English

EFL: English as a foreign language

ELF: English as a lingua franca

L1: First language

L2: Second language

NSE: Native speaker of English

NNSE: Non-native speaker of English

OED: Oxford English Dictionary

VOICE: Vienna-Oxford International Corpus of English

WCSE: Wellington Corpus of Spoken English

Tables

Figures

1 Introduction

> "All other things being equal, learners with big vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies, and there is some evidence to support the view that vocabulary skills make a significant contribution to almost all aspects of L2 proficiency."
> (Meara 1996: 37)

Vocabulary is an essential part of language proficiency and communicative competence. In fact, a growing body of evidence shows that a relationship exists between lexical knowledge and overall language ability (see for example Alderson 2005, Milton and Alexiou 2009, Milton *et al.* 2010, Stæhr 2008). Moreover, research suggests that a large amount of vocabulary is needed in order to function in English: the current research consensus is that knowledge of as much as 6,000–7,000 of the most frequent word families may be needed to enable comprehension of spoken discourse and 8,000–9,000 for comprehension of written discourse (Hu and Nation 2000, Nation 2006, Stæhr 2009, Laufer and Ravenhorst-Kalovski 2010). This represents a substantial learning challenge, and one that research indicates learners most often fail to reach (see Schmitt 2008: 332 for a review).

However, these findings are based on studies into the language usage of native speakers of English (NSE), which is not necessarily representative of the kind of language with which non-native speakers of English (NNSE) will engage. In fact, due to the status of English as a global lingua franca (ELF), Crystal (2006: 424–426) estimated that the number of non-native speakers who are able to communicate to a "useful level" in English were believed to outnumber the number of native speakers worldwide by approximately three to one already a decade ago. This means that today non-native speakers of English are more likely to use English with other non-native speakers of English than with native speakers of English (Jenkins, Cogo and Dewey 2011: 282). This calls into question the status of the native speaker of English as the only model for learners of English. Indeed, a growing number of scholars argue that the communicatively successful non-native speaker of English can represent a legitimate model for learners of English (Cook 1999, Seidlhofer 2004, Jenkins 2006, Mauranen 2011, Widdowson 2013).

Yet, a review of the literature reveals little research that has attempted to identify the number of words required to understand English in international contexts where it is spoken as a lingua franca. Indeed, only one researcher could be identified as having studied lexical coverage (i.e. the percentage of words that are accounted for in various corpora by particular word lists) in the use of English in international contexts: Gilner has focused on a core vocabulary (termed Dominant Vocabulary or DOVO) of up to approximately 1,200 word families (see her 2016 paper for a report on a collection of studies on the subject). These studies suggest that the amount of vocabulary used in contexts where English is spoken as a lingua franca (ELF) is lower than what has been found in chiefly monolingual and intranational English speaking contexts (see, for example, Nation 2006, and Schmitt *et al.* 2017 for an overview). If this is indeed the case, it would be reasonable to argue that the vocabulary size targets for learners of English as a foreign language (EFL) could be adjusted downwards, at least for those EFL learners whose aim it is to use English as a lingua franca in international contexts rather than in intranational environments where English is spoken predominantly as a native language.

Additionally, though speech is a primary medium of language, its relationship to vocabulary has generally been under-researched compared to written discourse (for an overview see Schmitt 2010: 9). Consequently, there is a large gap in our general understanding of vocabulary in spoken discourse, especially in ELF settings (see Gilner 2016: 28). This is the knowledge gap that the present study aims towards filling. Thus, the research questions that this study aims to answer are:

1. How much vocabulary is needed to understand English in international contexts where it is spoken as a lingua franca?
2. How does this compare to intranational contexts where English is spoken between native speakers of English?

The hypothesis of this study is that a smaller range of word families will be needed to understand spoken English in international contexts where it is used as a lingua franca compared to what has been found for intranational contexts where English is spoken amongst native speakers of English.

In order to answer the research questions, I will take a usage-based approach, which relies on the observation and analysis of language used in real-world contexts. Therefore, my analysis will be based on a corpus which aims to be representative of

naturally occurring spoken ELF, the Vienna-Oxford International Corpus of English (henceforth VOICE).

2 Theoretical framework

One of the key issues in teaching and learning English as a second language (henceforth L2) is determining the amount of vocabulary that needs to be known to enable communicative competence in language. This will often depend on the aims of the learner. For a learner wishing to achieve only a very basic degree of linguistic competence, enough to function in a limited range of situations, such as ordering a meal, or checking into a hotel, a very small amount of vocabulary knowledge might suffice. At the other extreme, an unrealistically ambitious learner might wish to master all of the existing words in the language, which, in addition to an unfathomable amount time and commitment, would require an understanding of how many words exist in the language.

According to Goulden, Nation and Read (1990), who based their estimates on Webster's Third International Dictionary (1963), the figure stood at around 58,000 word families at the time of the study, i.e. consisting of base forms, inflected forms and transparent derivatives (excluding proper names, compound words and abbreviations). However, since even well-educated, first language, adult speakers (henceforth L1) of English do not know all the words in the language, such a target would be unrealistic for a learner of the language. In fact, research (Goulden *et al.* 1990, Zechmeister *et al.* 1995, see also Schmitt 2010 for a review) indicates that, by adulthood, well-educated L1 speakers know somewhere in the region of 16,000–20,000 word families (excluding proper names, compound words and abbreviations).

These figures may be far too high to represent sensible vocabulary size targets for L2 learners of English, since it has been estimated that they are generally only able to achieve an average vocabulary size of around 2,000 of the most frequent word-families after approximately 1,000 hours of instruction (see Table 1). Although it appears that it is not impossible for L2 learners to achieve vocabulary sizes akin to a well-educated adult L1 speakers of English, this does not seem to be the norm (Nation and Waring 1997). Indeed, Laufer and Yano (2001: 549) assert that on average even proficient adult L2 learners of English have an English vocabulary size which is less than a quarter of the size of their L1 counterparts. Thus, it seems that

what would be needed is an understanding of vocabulary size targets based on what is required to be functional in specific contexts, such as ELF contexts versus native-speaker contexts.

Table 1: Estimates of English vocabulary size of second language learners[a]

Country and school	Vocabulary size	Hours of instruction[b]	Reference
China. English majors	4,000	1,800–2,400	Laufer 2001
Israel, high school graduates	3,500	1,500	Laufer 1998
Japan, EFL university	2,300	800–1,200	Barrow *et al.* 1999
Japan, EFL university	2,000	800–1,200	Shillaw 1995
Oman, EFL university	2,000	1,350+	Horst *et al.* 1998
Greece, high school	1,680	660	Milton and Meara 1998
Indonesia, EFL university	1,220	900	Nurweni and Read 1999
Germany, high school	1,200	400	Milton and Meara 1998
France, high school	1,000	400	Arnaud *et al.* 1985

Notes
[a] The table is taken from Schmitt, 2008: 332, slightly adapted.
[b] The data on hours of instruction was largely obtained by Laufer's (2000: 48) personal communication with colleagues from the respective countries.

Another problematic issue is how to define the concept of a word, because what should count as a word depends very much on the task at hand. For an essay writing assignment words may be understood in one of their simplest forms, consisting of a single letter or string of letters offset by orthographic boundaries, i.e. blank spaces or punctuation marks. For the study at hand, the definition of a word is necessarily more complex: since the aim of this study is to inform language pedagogy, the definition of a word needs to take into account the current scientific understanding of how words are represented and processed in the mind.

2.1 Terminology

Before looking at the principal complexities of defining a word for the purposes of linguistic analysis aimed at informing language pedagogy, it is useful to briefly introduce some of the relevant key terminology for readers not familiar with lexical corpus linguistic research and methodologies, such as the ones applied in the present

study. First, a definition of *corpus* (plural *corpora*): in linguistics, this term refers to a structured (digital) collection of language texts, selected according to specific linguistic and extralinguistic criteria, with the aim of being a representative sample of a particular language variety or genre (Sinclair 2005: 12).

A selection of the terms used to denote lexical counting units in quantitative corpus linguistics are *tokens*, *types*, *lemmas*, *flemmas* and *word families*. I provide here definitions for each of these terms based on a general review of the relevant literature and with particular reference to the definitions collected in *A glossary of corpus linguistics* (Hardie, Baker and McEnery 2006).

Firstly, *tokens* are a unit of measurement that refer to the total number of individual (lexical) items in a corpus. More specifically, *tokens* are generally defined as lexical items made up of a single letter (such as the indefinite article *a*) or a string of letters (such as the common noun *girl*) which are orthographically separated from other lexical items by a blank space. Depending on the format of the corpus or the settings of the software being used, the definition of a token may also include punctuation or morphemes, such as *n't* or *'s*. Instead, the number of *types* refers to the total number of uniquely spelt word forms in a corpus. Thus, a sentence such as: "The girl played the saxaphone." contains five *tokens* and four *types*, i.e. five lexical items, of which four are unique, since the word *the* is repeated twice. Hence, a corpus may contain, for example, one million *tokens*, though many of those are likely to reoccur in the corpus, so that there may only be 20,000 unique *types*.

Another frequently used lexical counting unit in corpus linguistics is the *lemma*. It is defined by Francis and Kučera (1982: 1) as "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling". In other words, it is the basic word form, such as the singular form of a noun or infinitive form of a verb (e.g. PLAY), which is conventionally used (for example, in dictionaries) to represent a set of semantically related lexical items that belong to the same major word class and which vary only in inflection. Hence, the lemma of the noun PLAY consists of the singular form *play* and the plural form *plays*, whilst the verb PLAY is a separate lemma comprising the words *play, plays, played and playing*.

The *flemma* (Nation 2016: XIII) is a larger lexical unit than the lemma and a smaller one than the word family. It consists of a headword and the inflected forms of different parts of speech. For example, the flemma for the headword PLAY (verb

and noun) includes *play*, *plays* (the present tense, third person verb form and the plural noun form), *played* (the past tense form and the past participle) and *playing* (all parts of speech).

A review of the relevant literature shows that the final term to be introduced here, *word families*, is applied more loosely than the others discussed above. Moreover, the term is not included in *A glossary of corpus linguistics* (Hardie, Baker and McEnery 2006), presumably because the authors did not consider it a term which pertains, strictly speaking, to the field of corpus linguistics. However, it is a lexical counting unit that has been used in three corpus-based studies that serve as a reference for the present study (i.e. Adolphs and Schmitt 2003, Nation 2006, Gilner 2016). The usefulness of the construct of *word family* as a lexical counting unit is discussed in more detail later in this section, but first I provide an overview of the main way in which it has been conceptualised.

Essentially, the term *word family* is generally used to refer to a group of semantically related words formed through inflectional and derivational affixation from the same basic word form. Thus, it is a larger lexical counting unit than the flemma. For example, the word family for the base word PLAY could include the words *playable, played, player, players, playful, playfully, playfulness, playing, plays, replay, replayed, replaying, replays and unplayable*. (For the fullest currently available word family lists, see those created by Nation, described in Nation and Webb 2011: 131–156 and available from http://www.laurenceanthony.net/ [Last accessed on 20 October 2017]).

2.2 The construct of *word* from a psycholinguistic perspective

Since the main purpose of this thesis is pedagogical in nature, in that it aims to inform vocabulary size targets, the construct of *word* should also take into account psycholinguistic aspects of vocabulary acquisition. In other words, what lexical unit should count as a word for purposes of analysis should, as far as is practically possible, be based on the current scientific understanding about how knowledge of vocabulary is represented, stored and processed in the minds of language learners and users. In a review article of corpus studies into vocabulary, Gardner (2007) identifies three main aspects that he argues should be considered from the psycholinguistic perspective when deciding on the lexical unit of measurement:

morphology, form-meaning variation and multi-word lexical units. I will discuss each of these in turn.

One of the key issues that needs to be considered is the extent to which learners with different language skills and backgrounds are able to make connections between morphologically related words. In psycholinguistics, there is a longstanding debate which centres around the balance between storage and processing of morphological information in the brain (see Gagné 2017 for a review). One of the main questions concerns the degree of morphological composition of words stored in the mind versus that processed online. Although the area has been intensively researched there is still no consensus among scholars, and there are competing theoretical approaches which fall along a cline between a balance of morphemic storage and morphemic processing. On the one end, it is posited that all words, including morphologically complex ones, are stored and accessed as whole units, without decomposed morphemic representation. At the other end of the spectrum, it is believed, instead, that words are represented and processed as morphemic units. There is also a mixed model which suggests that there may be a dual system in operation: one which stores and accesses some words as whole units, and another which stores and accesses the base morphemic unit, and then processes inflectional and derivational affixes online (Gagné 2017).

In this dual system, it is postulated that, for example, the frequency of morphologically complex word forms may affect how they are represented and processed in the mind. In fact, it is a well-established finding that words which occur more frequently in a language are recognised much more quickly than words which occur less frequently (see Schmid 2017: 3–8 for a recent review). Thus, it is assumed that through frequent exposure and use, words become entrenched in the mind as holistic chunks rendering recall and recognition of the word automated, so that compositional processing is no longer required.

Additionally, research indicates that a number of variables affect how readily individuals access and process morphological knowledge, including age, general language proficiency and morphological training (see Gardner 2007 for a review). With regards to age, research suggests that children have inflectional knowledge of English, when it is their first language, already during the first grade of school. However, competence in derivational morphology begins to develop later, around the fourth grade, when children are nine years old, and it continues to develop into

adolescence (Carlisle 2000, Tyler and Nagy 1989) and possibly beyond (Tyler and Nagy 1990).

Adult L2 learners of English also acquire inflectional knowledge before derivational (Schmitt and Meara 1997, Schmitt and Zimmerman 2002), which the researchers ascribe to the rule-based character of inflectional morphology in English compared to the less predictable nature of derivational morphology. Thus, regular inflections of noun and verb forms are likely to be learnt in the early stages of English L2 learning, so that, even at an elementary level of language learning, an English L2 learner who encounters the word *dog* and its inflected plural form *dogs* could be expected to recognise that they are semantically related. Instead, derivational morphology has been found to present mixed problems for language learners, with prefixes, such as *non-* or *pre-,* being generally more transparent than suffixes such as *-ment* (Nagy, Diakidoy and Anderson 1993). Moreover, though L2 learners of English have been found to be less sensitive to morphological structure than their native English-speaking peers (see Clahsen *et al.* 2010 for a review), neither native speakers of English nor advanced users of English can be assumed to have complete control of derivational affixes.

Bauer and Nation (1993) offer a comprehensive seven-tier model for addressing several of the morphological variables in relation to a possible cline in learner recognition (see Table 2). In selecting the criteria for ordering the affixes, the researchers consider how likely it is that a L2 learner of English would be able to recognise the base word when combined with inflectional and derivational affixes. Thus, the earlier levels on the seven-tier model represent the earlier stages of L2 acquisition of English, whilst the later levels represent the more advanced stages of L2 English language learning. These criteria include frequency, productivity, predictability and regularity. Frequent affixes (such as *-er* as in *play**er***) are highly generalised, so the researchers believe that it is more likely that they will be more easily recognised by an L2 learner of English. For this reason, such affixes are placed in the earlier levels of the seven-tier model. Productivity refers to the likelihood that an affix will be used to form a new word. For example, the affixes *-ly* and *-ness* are highly productive affixes, so they are placed in earlier levels than less productive ones, such as *en-*. Predictability concerns how transparent the meaning of an affix is likely to be, e.g. the suffix *-less* is deemed to be fairly transparent, so it is placed in an earlier level than a less transparent suffix such as *-ery*.

Table 2: Difficulty order of L2 English affixes (Bauer and Nation, 1993)

Level	Description
Level 1	*A different form is a different word. Capitalization is ignored.* No awareness of morphological relationships is assumed. This is a potentially useful lexical counting unit for words with multiple meanings, such as *bear*(s) (the animal) versus *bear*, *bore*, *borne* (meaning to carry).
Level 2	*Regularly inflected words are part of the same family. The inflectional categories are - plural, third person singular present tense, past tense, past participle, -ing, comparative, superlative, possessive.* This is what Nation calls a flemma (2016: XIII). At this level, it is assumed that a learner is able to recognise the relationship between regularly inflected forms.
Level 3	*The most frequent and regular derivational affixes: -able, -er, -ish,-less, -ly, -ness, -th (fourth), -y, non-, un- (unusual), all with restricted uses.* At this level, the ranking criteria of productivity, predictability and regularity are applied strictly.
Level 4	*Frequent and regular affixes: -al (coastal), -ation, -ess, -ful, -ism, -ist, -ity, -ise (-ize), -ment, -ous, in-, all with restricted uses*. At this level, orthographic regularity is prioritised over phonological criteria. This is because the researchers were creating a model for reading comprehension.
Level 5	*Infrequent but regular affixes: -age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom, officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery, trickery), -ese (Japanese, officialese), -esque (picturesque), -ette (usherette, roomette), -hood (childhood), -i (Israeli), -ian (phonetician, Johnsonian), -ite (Paisleyite, also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory), -ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise, discussion-wise), anti- (anti-inflation), ante- (anteroom), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage, enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify, subterranean), un- (untie, unburden).* These affixes may be easily recognised, but they do not add greatly to the number of derived forms that can be understood because they are infrequent.
Level 6	*Frequent but irregular affixes: -able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-.* Whist these affixes are frequent, they may not be very transparent to a learner because they cause allomorphy in their bases.
Level 7	*Classical roots and affixes, e.g., -ate, -ure, etc.* These occur only as bound roots, and according to the researchers need to be explicitly taught to learners.

Regularity of the written and spoken base form regards the degree to which the base form changes when the affix is added, and, thus, how easily it might be recognised when deconstructed. For example, *-ish* (as in greenish) comes in the earlier levels. Instead, base forms that change when the affix is removed, such as in the case of the word *sacrilegious*, come in the later levels. Regularity of function is also considered as when an affix is consistently attached to a certain word-class, such as *-ess*, which always attaches to nouns.

The researchers state that this model has no theoretical value, but they hope that it can be useful for a transparent operationalisation of the construct of the word family. However, as Gardner points out (2007: 247) since many affixes (e.g. -y, -ist -able) can occur at several different levels, this could become problematic because each instance of such an affixed word would need to be categorised on a case by case basis. Thus, Gardner (2007: 258–259) proposes a simpler model of three degrees of morphological complexity that should be considered to establish a psycholinguistically valid lexical counting unit taking into account general English language proficiency and age related developmental issues.

Gardner recommends the following (2007: 258-259): (1) Firstly, for younger children and learners with low general English proficiency, Gardner (ibid.) suggests that the base word plus regular inflections should be as a lexical counting unit in studies which aim at informing L2 English language pedagogy. That is to say, the lemma form excluding irregular inflections. (2) Secondly, for older children and adolescents, as well as for learners with intermediate general English proficiency, he recommends using the base forms plus both regular and irregular inflections, as well as derivational prefixes. In other words, the lemma form plus prefixed forms. (3) Thirdly, for adults and learners with high general English proficiency he suggests using the word family, i.e. base forms plus both regular and irregular inflectional and derivational affixes.

Another key issue that needs to be considered when deciding what should count as a word for pedagogical purposes is related to whether to treat words that have the same form but multiple meanings as single or separate lemmas. This is the case for homographs, such as the word *bank*, which can, for example, refer to either a financial institute where people can keep their money or to the side of a river.

Therefore, a psycholinguistically valid operationalisation of the construct of a word in a study such as this should ideally also take into account whether L2

language learners would perceive such words as one word or as separate words, the intended meaning of which is generally made clear by the context of the word. In a machine based count of lexical items, such words may be indistinguishable and, so, counted within the same category, but it is unclear whether a language learner would make any connection between these words when encountered in their respective contexts. Thus, counting such polysemous homographs as a single word could lead to an underestimation of the learning burden for L2 learners of English.

Moreover, corpus-based studies have revealed that higher frequency vocabulary, that is lexical items that are generally used more frequently in a language, tend to be more polysemous than lower frequency words (Ravin and Leacock 2000: 1). Additionally, the less technical and more general the semantic field of a word, the greater the variation in form and meaning tends to be. Hence studies of general language, unspecialised corpora, like the VOICE corpus used in this study, are likely to contain a higher number of polysemous lexical items (Gardner 2007: 253). With this in mind, Gardner (2007) cautions that adjustments need to be made where possible, and where they are not possible, the analysis of data and the reporting of results need to be carried out with awareness and transparency regarding possible limitations and threats to validity.

The final issue that Gardner (2007) points out should be kept in mind for a psycholinguistically valid conceptualisation of a lexical counting unit are fixed or semi-fixed multi-word items, which form semantically inseparable units: these include open compounds (*air conditioning*), phrasal verbs (*put up with*), idioms (*rock the boat*), fixed expressions (*good afternoon*), and prefabs (*the point is*). There is substantial evidence that native speakers of English (for a review see Schmid 2017: 7) store and access these (semi-)holistically, without the need for online composition. However, for non-native speakers the evidence is mixed, with only proficient users showing signs of some or partial holistic representation and retrieval (for a review see Conklin and Schmitt 2012). However, such formulaic language can be difficult to identify and process electronically.

Taking into account both the psycholinguistic considerations discussed above and the practical constraints of doing a machine-based analysis of VOICE, I have chosen, in this study, to provide lexical coverage figures for three levels of morphology: the *word type*, the *flemma* and the *word family*. I have made accommodations for polysemous words in the count of the word families by placing

them in separate word families. However, multi-word units have not be taken into account in this study due to constraints in time and resources.

2.3 Previous research into lexical coverage in spoken discourse

Since the aim of this study is to determine the amount of vocabulary required to predict comprehension of spoken discourse in contexts where English is spoken as a lingua franca, I review previous related research findings in this section.

Research has documented strong correlations between second language skills and lexical knowledge (see for example Bonk 2000, Albrechtsen *et al.* 2008, Alderson 2005, Laufer 1992, Laufer and Goldstein 2004, Stæhr 2008, Wang 2017). Such research indicates that lexical competence is central to second language performance, and indeed, based on his findings, Alderson (2005: 88) concludes that "language ability is to quite a large extent a function of vocabulary size". Second language (L2) learners also typically recognise the importance of vocabulary for effective comprehension and expression (Read 2004: 146). For example, Simon and Taverniers (2011: 912) found that English as a foreign language (EFL) learners believed that difficulties with vocabulary were more likely to cause communication breakdown than both grammar and pronunciation errors.

Given that research indicates that lexical competence is essential for general language competence, it is important to establish the amount of lexical knowledge needed to be functional in a wide variety of contexts. This information is useful for L2 teachers and learners to set vocabulary size targets. One of the key questions regards how many of the words in a piece of spoken or written discourse a language user needs to know to support understanding. According to a recent review by Schmitt *et al.* (2017: 214), the current consensus amongst researchers is that knowledge of at least 98% of the running words (i.e. words which follow each other consecutively) is a significant predictor of reading and listening success (see for example Carver 1994, Hu and Nation 2000), though also a minimal level of 95% (Laufer 1989, 1992; Laufer and Ravenhorst-Kalovski 2010, van Zeeland and Schmitt 2012). This percentage of known words in a text is referred to as "lexical coverage" in the literature.

Laufer (2010: 16) also defines and explains another important term related to the relationship between lexical knowledge and successful language use: the term "sight vocabulary", which refers to words whose meaning is so familiar that they can

be understood without relying on the context to infer their meaning. This makes it possible to decode these words quickly, which frees cognitive effort for the higher-level processes needed for understanding the contents and implications of discourse. Hence, the larger the sight vocabulary, the higher the lexical coverage.

The amount of lexical coverage needed depends on the desired or required degree of comprehension, as well of the demands of the specific task or context. Thus, the lexical coverage figures cited above have depended on how the researchers have operationalised the construct of "adequate" comprehension, as there is not one definitive and established understanding of this concept. Additionally, it is also important to note that even if all the words in a given text are known by the reader or listener, this does not necessarily guarantee that the text will be understood. This is because knowledge of vocabulary is only one of the factors involved in language use. Others include, for example, grammatical knowledge, background and world knowledge and skill in language use, including the use of compensatory strategies.

Nonetheless, since lexical knowledge has been found to be one of the main predictors of successful language use, several studies have attempted to establish lexical thresholds for various types of language performance. The relationship between receptive vocabulary knowledge and reading comprehension, in particular, has been widely researched, and strong correlations have been found, ranging from between 0.40 and 0.85 (Bonk 2000, Henriksen *et al.* 2004, Laufer and Ravenhorst-Kalovski 2010, Qian 1999, 2002; Stæhr 2008, Schmitt *et al.* 2011). These can be considered strong correlations because as mentioned above vocabulary knowledge is only one of the factors involved in successful language use, and yet it has been found to be such a significant predictor of comprehension.

The first attempt to relate L2 knowledge of vocabulary to success in reading comprehension was made by Laufer (1989). The lexical coverage of the learners (n=100 first year Israeli university students whose L1s were Hebrew and Arabic) was determined by the learners' self report: they underlined unknown words in an academic text. Reading comprehension was measured with a reading comprehension test and adequate comprehension was set at a score of 55%. It was found that with a threshold of 95% lexical coverage most participants achieved this score. However, as Laufer notes (2010: 17) "most educators [...] would probably not be satisfied with such a low [comprehension] score." Hence, greater lexical coverage would probably be desirable.

Another study investigated how the reading comprehension of learners of English was affected by lexical coverage of a fiction text (Hu and Nation 2000). This was determined by substituting low-frequency words in the text with nonsense words to ensure that the words were unknown to all participants. Four levels of lexical coverage were set: 80%, 90%, 95% and 100%. Hence, at 80% coverage one in five words was a nonsense word, at 90% the figure was one in ten, at 95% it was one in 20 words, and at 100% none of the words were replaced with nonsense words. The other words in the text belonged to the 2,000 most frequent words in English. The participants (n=66 adults attending a pre-university course) were tested for their knowledge of the 2,000 most frequent words to ensure that the learners would not have difficulty with the vocabulary in the text, apart from the nonsense words. The participants reading comprehension was then tested by means of multiple choice questions and by a written cued recall of the text. The level of "adequate" comprehension was set at the level where the majority of the participants in the 100% group (i.e. those whose text contained no nonsense words) were judged to have understood the text: i.e. they achieved a score of 12 out of 14 on the multiple-choice questions, and 70 out of 124 on the written recall test. The multiple-choice test was trialled with native speakers of English before being used in the study with the L2 learners of English.

It was found that with 80% lexical coverage (i.e. one nonsense word out of every five, or 20 out of 100) none of the participants gained adequate comprehension. At 90% and 95% lexical coverage, some attained adequate comprehension, but most did not. At 100% lexical coverage, most participants reached the threshold set for adequate comprehension. The researchers applied a regression model to the data to determine a reasonable fit, and it was calculated that 98% lexical coverage (i.e. one unknown word in 50) would be a good predictor of adequate, unassisted comprehension of the text for the majority of learners.

This is in line with Carver's findings (1994: 432) with native speakers of English: 219 students in Grades 3, 4, 5, and 6 and 60 graduate students in America. He investigated the relationship between the relative difficulty of two written texts and the number of unknown words in the texts: one factual and another fictional, which had been sampled from the school curriculum and library books. The relative difficulty of the texts was determined from the difference between the difficulty of the material and the reading ability of the students in grade equivalent units. The

number of unknown words was identified by having the participants underline words for which they did not know the meaning. The researcher found that when no words are unknown in the reading material then it tends to be regarded as relatively easy, whereas when two per cent of the words are unknown, then the reading material is considered comparatively difficult. At around one per cent of unknown words, the level of reading difficultly is roughly equal to the ability level of the individual. Based on his findings, Carver notes that even 98% lexical coverage does not render reading comprehension easy.

Schmitt, Jiang and Grabe (2011) also explored the relationship between lexical coverage and reading comprehension. Their mixed L2 participants (n=661) took a vocabulary test for words taken from two texts, and then completed a reading comprehension test for the texts. The researchers found a relatively linear relationship between the percentage of vocabulary known and the degree of reading comprehension within the coverage range of 90% and 100%, with no evidence of an absolute lexical threshold where comprehension increased greatly. Hence, the researchers concluded that the necessary lexical coverage depends on the required degree of comprehension. Based on their findings, if 60% comprehension is required then a lexical coverage of 98% would be required.

Much less is know about known about the lexical coverage required to predict listening comprehension. Stæhr (2009) indirectly measured the effect of lexical coverage on listening comprehension. The listening comprehension of his Danish participants (n=115) was assessed with the Cambridge Certificate of Proficiency in English (CPE), their vocabulary size was measured with the updated version of Nation's Vocabulary Levels Test (VLT) developed by Schmitt, Schmitt and Clapham (2001), and their depth of vocabulary knowledge with a version of the Word Associates Test (Read 1993, 1998).

To measure the lexical coverage, the participants' scores on the Vocabulary Levels Test were matched to the vocabulary frequency profiles of the listening passages, which were obtained with the Vocabulary Profiler on the Compleat Lexical Tutor website (Cobb 2000). For example, participants who mastered the 5,000 VLT level were assumed to have 98% lexical coverage of the listening passages, whilst at the 3,000 VLT level participants were assumed to have 94% lexical coverage of the texts. With this method, Stæhr (2009) indirectly tested the effects of lexical coverage

on listening comprehension, and found a linear relationship between them, confirming Schmitt, Jiang and Grabe's findings (2011) for reading comprehension.

Stæhr found a significant correlation between listening comprehension and both size and depth of vocabulary knowledge (at 0.70 and 0.65, respectively). Additionally, it was found that 98% lexical coverage led to a mean listening comprehension score of 73%, whilst at 94% lexical coverage, the comprehension score was found to be significantly lower, at 59%. Stæhr (2009) concluded that the lexical coverage threshold will inevitably depend on the level of comprehension required.

A more recent study by van Zeeland and Schmitt (2012) has tested the effect of lexical coverage and listening comprehension more directly. The researchers had their mixed L1 participants (n=76: 36 native and 40 non-native speakers of English) listen to four anecdotes. Varying percentages of words were replaced with nonsense words in the passages (0, 2, 5 and 10 per cent), so that they contained respectively 100, 98, 95 and 90 per cent of known words. Participants' comprehension was tested with ten multiple choice questions about factual information, and it was found that greater lexical coverage led to better comprehension: when participants knew 100% of the words in the story they obtained a mean score of 9.62 out of 10 on the listening comprehension test; at 98% lexical coverage, the mean comprehension score was 8.22; at 95% it was 7.65 and at 90% it was 7.35. Though listening comprehension was still relatively good with 90% lexical coverage, the individual variation at this level was high. Thus, the researchers concluded that 95% lexical coverage was a more suitable threshold for adequate comprehension because, at this level, participants performed more consistently.

This understanding of the relationship between lexical coverage and successful listening and reading comprehension prompts another question: how much vocabulary does it take to achieve this level of lexical coverage of written and spoken discourse? In general, according to a review by Schmitt *et al.* (2017: 214), the amount of enquiry which can inform vocabulary size targets is limited and would need to be expanded to provide useful information to language learners and teachers. One of the main goals of language learners and users is to gain oral communicative competence, so information about how much vocabulary is needed to support this language skill is useful.

A study which seeks to provide such information is Adolphs and Schmitt (2003). The study was based on an analysis of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE). CANCODE is a five-million-word corpus of spontaneous conversations and interaction in English amongst people from all segments of society in Britain and Ireland. It was collected between 1995 and 2002. The researchers also supplemented this an analysis of a section of the spoken component of the British National Corpus (BNC): they examined approximately four-and-a-half million words from the BNC, made up of conversational data, as well as other everyday language use, such as meetings, lectures and sermons.

The lexical unit used in this study was word families and the construct included all inflected forms and suffixed derivatives, but not prefixed ones. Open compounds (e.g. Prime Minister) were counted separately, and homographs were not distinguished and were counted under the same word family. Finally, the researchers included backchannels (such as *eh* and *uh huh)* grouped under a single category, since Biber *et al.* (1999) have found that these carry a great deal of meaning and are a common feature of spoken discourse.

For the comparative analysis with the BNC, the researchers used individual word types instead of word families as the basic unit of their calculation. In order to calculate the lexical coverage, the researchers entered the list of word families/individual words and their frequency of occurrence into a spreadsheet. Then they divided the total number of words occurring at various levels (e.g. the most frequent 2,000 or 3,000 word families, or 5,000 individual words) by the total number of words in the corpus to arrive at a percentage of text coverage.

The study found that the most frequent 2,000 word families offer around 95% coverage of their corpus, whilst 96% coverage was achieved with the most frequent 3,000 word families or 5,000 individual words. The only other directly comparable previous study, carried out by Schonell, Meddleton and Shaw (1956) with Australian workers, found that 2,000 word families offered a coverage of 98–99%. Instead, Adolphs and Schmitt's (2003) findings indicate that learners of English would need to acquire a considerably higher number of words to reach the same level of coverage.

Though this study has important implications for informing English language instruction, the operationalisation of the construct of the lexical counting units used in the study may have certain validity problems from a psycholinguistic point of

view: i.e. derivations formed with prefixes were counted separately. According to Gardner's review of the literature, research indicates that they tend to be learnt sooner than derivational suffixes (2007: 258–59), so this treatment may not have been optimal. Another potential limitation of the study, which the researchers themselves note (Adolphs and Schmitt 2003: 432), is that it is not clear from research that 95–96% coverage is adequate to enable functioning in a wide variety of contexts where English is spoken. Therefore, it would probably have been useful to provide coverage figures for higher levels also, e.g. 98%.

Finally, Adolphs and Schmitt (2003: 430–32) argue that the CANCODE corpus is "likely to be more representative of the kind of spoken discourse that the typical [...] L2 learner would be in contact with..." However, since the corpus is only representative, at best, of British and Irish native-speaker discourse, this completely disregards the fact that nowadays non-native speakers of English are more likely to use English with other non-native speakers than native speakers of English (Crystal 2006: 424–426, Jenkins, Cogo and Dewey 2011: 282).

The Adolphs and Schmitt (2003) study only checked the lexical coverage offered by the word families within CANCODE of the corpus itself. Instead, in a study which attempted to establish what lexical coverage a more generally representative sample English would give of a variety of written and spoken text types, Nation (2006) concluded that knowledge of the most frequent 4,000 word families (plus proper nouns) provides around 95% lexical coverage of a wide range of authentic written texts, and with about 8,000–9,000 word families (plus proper nouns) a coverage of 98% can be reached. For general (non-technical) spoken discourse, about 3,000 word families (plus proper nouns) allow for over 95% lexical coverage, and with about 6,000–7,000 word families (plus proper nouns) 98% coverage can be gained.

This study (Nation 2006) is one of the most recent and widely cited investigations into the number of words needed to reach various levels of lexical coverage. However, the study was rather limited in terms of its representativeness. Nation carried out his investigation by profiling various genres against fourteen 1,000 word-family lists. These lists "are sequenced largely according to their range, dispersion and frequency in the 10 million spoken section of the BNC" (Nation 2006: 80). Therefore, the reference word list was a sample of only general British English.

Nation (2006) used the word-family lists to separately frequency profile a text collection of five novels: *Lord Jim* by Joseph Conrad (originally published in 1900), *Lady Chatterley's Lover* by D. H. Lawrence (originally published in 1928), *The Turn of the Screw* by Henry James (originally published in 1898), *The Great Gatsby* by F. Scott Fitzgerald (originally published in 1925), and *Tono-Bungay* by H. G. Wells (originally published in 1909); a collection of parallel newspaper corpora (forty-four 2,000-token collections of news articles from the *Lancaster-Oslo/Bergen Corpus, the Freiburg–LOB Corpus of British English, the Brown, Frown*, and *Kolaphur* corpora); unscripted spoken English (two 100,000-token sections of the Wellington Corpus of Spoken English: radio call-ins and conversation); and finally the movie *Shrek* (released in 2001). Therefore, also the corpora against which the BNC-word-lists were compared are fairly limited in terms of representativeness. For example, the fiction is limited to out-of-copyright novels and does not contain any contemporary literature, and the unscripted spoken corpus is only representative, at best, of English as it is spoken in New Zealand.

Another somewhat relevant study was conducted by Gilner (2016). The paper reports on the findings of several previous studies aimed at identifying a core vocabulary for English used in localised and globalised settings. The researcher compared word families across a variety of corpora: the International Corpus of English (ICE), the 26 English Varieties corpus (26EV), the Vienna-Oxford International Corpus of English (VOICE) and the Corpus of English as a Lingua Franca in Academic Settings (ELFA).

ICE aims to be a representative sample of English varieties spoken within national contexts around the world where English is either used as a first language or an official additional language. In particular, the researcher examined the national subcomponents of the ICE corpora that were freely available at the time of the study: those for Canada, East Africa, Hong Kong, India, Jamaica, the Philippines, and Singapore. Each of these subcorpora comprise a one-million-word sample: 60% spoken discourse and 40% written discourse. The varieties represented in the 26EV corpus of 15 million words of written discourse were those of English used in Australia, the Bahamas, Belize, Bermuda, Cameroon, Canada, Fiji, India, Ireland, Jamaica, Kenya, Liberia, Malawi, Malaysia, Myanmar, New Zealand, Nigeria, Pakistan, the Philippines, Singapore, South Africa, Sri Lanka, Trinidad and Tobago, the United Kingdom, the United States, and Uganda.

The VOICE corpus (which is used also in the current study, so it is more fully described in the methodology section of this paper) is a one-million-word sample which aims to be a representative sample of general English spoken as a lingua franca in international settings. Finally, ELFA is also a one-million-word sample of English spoken as a lingua franca in the specialised context of academia.

The researcher identified a relatively small set of high frequency word families, the so-called ICE-CORE of 1,206 words families, that make up a significantly large proportion (87.9%) of the word families used across the sub-corpora of the ICE corpus. The ICE-CORE provided around 82.7% lexical coverage of 26EV. Based on these findings, the researcher hypothesises that a preferred *dominant vocabulary* (DOVO) is used in the discourse of English language users around the world.

Interestingly, the researcher found that the lexical coverage of the ICE-CORE for the ELF corpora was higher: at 90.24% and 92.67% for ELFA and VOICE respectively. This comparison between the intra-cultural (localised) communities and inter-cultural (international) communities led Gilner (2007: 48) to conclude that a greater reliance on the DOVO can be attributed to a convergence strategy used by ELF speakers to "bridge linguacultural divides". This finding supports the hypothesis of the present study that a smaller number of word families will be needed also for higher lexical coverage thresholds of 95% and 98%, which research indicates may necessary to function adequately across a wide range of contexts (see Schmitt *et al.* 2017: 214 for a recent review).

2.4 English as a Lingua Franca

The main aim of this study is to complement research which seeks to identify the amount of vocabulary needed to function in spoken English. As discussed in the previous section of this chapter, to date, research that has specifically attempted to set a figure for this has focused on data samples of English spoken by native speakers of the language (see Adolphs and Schmitt 2003, Nation 2006). Due to the current status of English as a global lingua franca, this data can provide, at best, a very partial account of the amount of vocabulary used in English as it is spoken around the world today.

In fact, English is currently the predominant language of international communication in business, science, research and politics. Accurate figures for the number of English speakers worldwide are difficult, if not impossible to obtain. One

of the reasons for this is that defining who qualifies as a speaker of a language is unavoidably ambiguous. That said, based on a review of resources from international organisations, linguistic surveys and individual authors, Crystal (2006: 424) made an informed estimate that approximately 1.5 billion people around the globe are conversant in English. Of those, roughly 400 million speak English as a first language, 400 million as a second language and between 600 to 700 million as a foreign language. Thus, already a decade ago approximately one fourth of the world population were able to communicate in English to a "useful level" (ibid: 425), with non-native speakers of English outnumbering native speakers by a ratio of almost three to one.

According to *Britannica Academic* (2018), the number of speakers of English worldwide now stands at some two billion. Hence, it is easy to imagine that the ratio of non-native to native speakers of English must surely have grown over the last decade, especially considering that the population growth of countries where English is used extensively as a second language, such as India, is considerably higher than that of countries where it is mainly used as a first language, such as the United States of America.

In the literature, a common starting point for making such comparisons is Kachru's Three Circles of Englishes model (1985): the Inner Circle, the Outer Circle and the Expanding Circle. This is one of the most influential models for grouping varieties of English around the world, though its limitations have been debated by a number of influential scholars including Kachru himself. It is, nonetheless, a useful stepping stone for discussing the prevailing ideologies that are attached to the English language, as well as for positioning the present study. For this reason, I will give a brief overview of the model and discuss some of its implications for users of English worldwide.

The circles represent "the type of spread, the patterns of acquisition and the functional domains in which English is used across cultures and languages" (Kachru 1985, 12). The Inner Circle refers to the nations where English is the dominant first language of the majority of the population. This includes the United States of America (316.5), the United Kingdom (64.1), Canada (35.2), Australia (23.1) and New Zealand (4.5). (The population sizes as of 2013 for these and the following countries are indicated in brackets in millions, unless otherwise stated. The source is the World Bank 2016.) These have traditionally been seen as the "norm-providing"

varieties of English, with particularly the British variety, and more recently the American variety, holding positions of prestige within English language teaching (Kachru 1985: 16).

The Outer Circle refers to the countries which were affected by the early spread of English mainly through colonisation, and in which English is now a dominant institutionalised language, playing an important intranational as well as international language role in a multilingual environment. Countries included in the Outer Circle are, for example, India (1.3 billion), Nigeria (173.6), Kenya (44.4), Malaysia (29.7), Singapore (5.4), amongst others. The English used in these countries has been referred to as "norm-developing" in that they are both "endonormative and exonormative" (Kachru 1985: 17).

The Expanding Circle indicates the rest of the world, where English has not been introduced through colonisation, but is the foreign language of choice for international communication. The current global status of English means that it has become the most studied foreign language by children around the world. For example, according to the European Commission (2016), English is still the most commonly studied foreign language at lower secondary level across the member states of the European Union: with 96.7% of pupils learning it, far ahead of French (34.1%), German (22.1%) and Spanish (12.2%). Countries in the Expanding Circle have been described as being "norm-dependent" and "exonormative" (Kachru 1985: 17) in that they have traditionally been viewed as aspiring to conform to the standards of the prestige varieties of the Inner Circle. These include, for example, China (1.4 billion), Russia (143.5), Japan (127.3), European countries apart from the United Kingdom and Southern American countries.

As Kachru (1985: 14–15) pointed out already three decades ago, this internationalisation of English brings to the language a "unique cultural pluralism, and a linguistic heterogeneity and diversity" the extent of which is unprecedented in the history of humankind. The sheer magnitude of this global diffusion has important implications for the description, codification and standardisation of English. Thus, the native speakers of the language, who have become a minority, have "lost the exclusive prerogative to control its standardization" (ibid: 30). This view is echoed by other scholars, for example, Widdowson (1994) too has questioned the "ownership" of English, as he puts it.

A quarter of a century after Kachru called for a reconceptualization of the localised varieties of the Outer Circle, it is now generally acknowledged that varieties that have developed endonormatively, such as Indian English, have gained legitimacy as Englishes in their own right (Seidlhofer 2011: 3). English used as an international lingua franca, however, is different. It is this international rather than intranational English language usage that the Three Circles model fails to capture, according to Seidlhofer (2011: 4). This use involves people across the three concentric circles, who use the language as a convenient common tool for communication, as a lingua franca.

English used as an international lingua franca (ELF) has been defined by scholars in two main ways. The first sees it as an additionally acquired language system used as a means of communication between people who do not share a common first language (Seidlhofer 2001: 146). An important aspect of this perspective of ELF is that it does not exclude native speakers of English (NSE) from ELF communication. Instead, it sees NSE as also having to acquire the ability to communicate effectively in contexts where English is used as a lingua franca (Jenkins, Cogo and Dewey 2011: 283).

This view differs from another earlier held view that, instead, saw English used as a lingua franca as "a 'contact language' between persons who share neither a common native tongue nor a common (national) culture and for whom English is the chosen *foreign* language of communication" (in Firth 1996: 240 the emphasis is in original). This second characterisation of English used as a lingua franca positions it as foreign language usage. However, most researchers of ELF today view it as fundamentally distinct from English as a foreign language. English as a foreign language (EFL) is generally seen as a deficient form of English aspiring to native English language competence. However, native-speaker norms may be irrelevant in international business and academia where English is used as a lingua franca. Jenkins, Cogo and Dewey describe ELF as "freed from the standardizing constraints of a set of norms" (2011: 291). In fact, as Smith (1978: 11) wrote almost four decades ago: "Today few people are willing to sound like native English speakers or to identify with their culture as is typically required in the second language situations".

As Nelson asserts, native speakers of English are a rare sight in most international interactions in English and many "may never have had the dubious

good fortune even to have met a native speaker" (1995: 276). This calls into question the status of the native-speaker of English as the only model for learners of English. Indeed, a growing number of scholars argue that the communicatively successful non-native speaker of English can represent a legitimate model for learners of English (Cook 1999, Jenkins 2006, Mauranen 2011, Seidlhofer 2004, Widdowson 2013). Cook (1999) advocates that "L2 users" should be reconfigured as successful intercultural speakers, instead of "failed" and "deficient" speakers. Mauranen (2011: 164–5) also suggests that the study of English as it is used in international contexts can be useful in informing foreign language teaching by showing how English actually works in real life beyond the classroom.

It is, in fact, this usage-based, descriptive rather than prescriptive approach that is the guiding principle of the current study. One of the basic assumptions of this study is that the VOICE corpus data is representative of generally successful communication. This assumption is based on the fact that the data analysed in this study is collected from "experienced ELF speakers" (VOICE - Corpus Description 2013), with one of the sampling criteria of the corpus being "[s]elf-selected participation (i.e. the speakers decided for themselves that they are capable of using ELF to accomplish specific participant roles in the speech event they are taking part in)" (VOICE - Corpus information 2013).

3 Materials and methods

3.1 Material

To determine how many words are typically used in contexts where English is spoken as a lingua franca, I used the largest general corpus of ELF currently available: the VOICE corpus (corpus version: VOICE POS XML 2.0). This corpus aims to be generally representative of ELF, especially as it is spoken within Europe. VOICE is a one-million-word sample of ELF, based on approximately 110 hours of audio-recordings of 151 naturally-occurring, non-scripted, face-to-face interactions. The recordings were carried out between July 2001 and November 2007, and are made up of complete speech events from a variety of domains (educational [23%], leisure [17%], professional [60%]) and speech event types (conversations, interviews, meetings, panels, press conferences, question-answer sessions, seminar

discussions, service encounters, working group discussions and workshop discussions).

VOICE includes approximately 1,250 speakers (753 identified individuals) from 49, mainly European, language backgrounds. The largest group of first language speakers represented in the corpus are L1 German (25%); English (7%); Dutch and Spanish (6%); French (5%); Finnish and Italian (4% each); Danish, Polish and Norwegian (3% each); and the remaining languages each make up two per cent or less of the speakers in the corpus. Although native speakers of English make up only 7% of the participants, they were present in 40% (*n*=61) out of 151 of the speech events (based on my own analysis of the data). The gender distribution of participants is roughly equal (51% female and 49% male), and the age distribution ranges from 17 to over 50. The speakers are self-selected as being capable of carrying out their communicative purpose in spoken English within the given domains and speech event types. (VOICE - Corpus Information 2013)

VOICE was compiled in accordance with generally accepted ethical and scientific principles: the data was collected with the informed consent of the participants and it was anonymised to protect their personal identities (VOICE - Corpus Information 2013).

3.2 Methods

3.2.1 Lexical coverage

The lexical counting units used in this study are the word type, the flemma and the word family. The construct of the flemma is based on the definition provided by Nation (2016: 26) and includes the inflected forms of different parts of speech under the same headword. Hence, for example, the inflected forms of both the verb and noun forms of LOOK are included under the same headword, but derived forms, such as *looker* and its inflected form *lookers* are counted under a separate headword. Instead, the construct of word family includes all inflected forms and derivatives formed with affixes up to level 6 on Bauer and Nation's (1993) seven-tier model of morphological affixation (see Table 2 in section 2.2 above for details). The compilation of the word family list for VOICE is based on the twenty-five frequency-ranked 1,000-word family lists created by Nation and available from

http://www.laurenceanthony.net/software/antwordprofiler/ (Last accessed on 4 February 2018).

The ranking of Nation's twenty-five 1,000-word family lists is based on range, distribution and frequency data from the British National Corpus (henceforth BNC) and the Corpus of Contemporary American English (henceforth COCA). However, in order to achieve a more balanced representation of words that are common in spoken English, a special procedure was used for the ranking of the first two 1,000-word family lists: these were ranked according to the range and frequency data of a ten-million-token sub-corpus made up of six million words of spoken English and four million words of written English. Additional adjustments included categorising the numbers (e.g. ten, thirty) and the days of the week in the first 1,000-word group and the months of the year in the second 1,000-word group even if their frequency did not always justify this. This was done in order to create a ranked list of words that would be suitable for course and material design for the teaching of English as a second language (see Nation and Webb 2011: 131–156 for a full description of the lists).

The word lists (henceforth BNC/COCA word lists) were generated with a computer programme which does not distinguish between polysemous homographs such as *book* (as in printed pages) and *book* (as arranging for the use of a table, hotel room, etc.). An attempt was made to deal with potential frequency ranking differences between homographs by placing them under different headwords, so that, for example, the noun form of *book* and *books* was placed in one word family, whilst the verb form of *book*, *books*, *booked*, and *booking* were placed in another. This does not completely account for possible frequency ranking differences between homographs, but it goes some way towards doing so. Table 3 shows a list of the homographs that were identified in VOICE, along with the frequency grouping of each of the pairs of homographs by word class.

Proper nouns were grouped as a discrete category irrespective of their frequency. The proper nouns category included words normally written with a capital letter which are (anonymised) names of people, places, institutions, etc., for example Ben, Erasmus, Klimt, Microsoft, Saab, Spider-Man, Yemen, and so on. Furthermore, Open compounds (e.g. *Prime Minister*) and multi-word units were counted separately.

Table 3. Identified homographs in VOICE by word class and frequency group

Headword	Word Class	Frequency Group	Word Class	Frequency Group
appropriate	adj, adv or noun	3rd 1,000	verb	5th 1,000
board	noun	1st 1,000	verb	4th 1,000
book	noun	1st 1,000	verb	4th 1,000
bound	verb (pp *bind*)	2nd 1,000	noun or base verb	4th 1,000
box	noun	1st 1,000	verb	4th 1,000
can	modal verb	1st 1,000	noun	2nd 1,000
chair	noun	1st 1,000	verb	6th 1,000
fair	adj	1st 1,000	noun	8th 1,000
fast	adj or adv	1st 1,000	verb	8th 1,000
fine	adj or adv	1st 1,000	verb	4th 1,000
firm	adj or adv	2nd 1,000	noun or verb	2nd 1,000
flat	adj	1st 1,000	noun	2nd 1,000
frank	adj or adv	2nd 1,000	proper noun	Proper noun
good	all other	1st 1,000	plural noun	3rd 1,000
kid	noun	1st 1,000	verb	6th 1,000
last	adj, adv or noun	1st 1,000	verb	2nd 1,000
lean	verb	2nd 1,000	adj	10th 1,000
long	adj or adv	1st 1,000	verb	4th 1,000
March	proper noun	2nd 1,000	verb	4th 1,000
mean/ing	verb or noun	1st 1,000	adj	1st 1,000
mine	pronoun	1st 1,000	noun or verb	2nd 1,000
minute	noun	1st 1,000	verb	19th 1,000
moderate	adj	3rd 1,000	verb	9th 1,000
patient	adj	2nd 1,000	noun	3rd 1,000
prospect	adj or noun	3rd 1,000	verb	9th 1,000
second	adj or adv	1st 1,000	noun	2nd 1,000
shorts	plural noun	5th 1,000	all other	1st 1,000
sound	noun or verb	1st 1,000	adj	7th 1,000
state	verb	1st 1,000	adj or noun	2nd 1,000
strand	noun	4th 1,000	verb	6th 1,000
stuff	noun	1st 1,000	verb	4th 1,000
subject	noun	1st 1,000	verb	4th 1,000
type	noun	1st 1,000	verb	5th 1,000
well	adv	1st 1,000	noun or verb	4th 1,000

Interjections and response particles were retained in the tokens for analysis since they have been found to be an important feature of spoken discourse, as they carry a great deal of meaning (Biber *et al.* 1999). However, hesitation markers and fillers (i.e. *eh*, *er* and *erm*) were excluded because they were considered to be more similar to pauses than to words. The interjections and response particles were assigned to 17 headwords according to the meaning that they convey. The descriptions of the categories are the following:

3. *huh* used as a tag-question;
4. *yay*, *yipee*, *whoohoo*, *mm* used as an exclamation to express joy or enthusiasm were grouped under the headword *yay*;
5. *hm, hmm, haeh* used to express doubt, disbelief or hesitation were grouped under the headword *haeh*;
6. *gosh, ah, oh, wow, poah* used to express astonishment or surprise were grouped under the headword *wow*;
7. *oops* used to express an apology;
8. *ooph* used to express exhaustion;
9. *ts, pf* used to express dismissal or contempt were grouped under the headword *ts*;
10. *ouch, ow* used to express pain were grouped under the headword *ouch*;
11. *sh, psh* used to request silence were grouped under the headword *sh*;
12. *oh-oh, uh* used to express the anticipation of trouble were grouped under the headword *oh-oh*;
13. *ur, yuck* used to express disgust were grouped under the headword *yuck*;
14. *oow* used to express pity or disappointment;
15. *blah* used to express lack of interest for something;
16. *gee* used to express annoyance;
17. *aha, mhm, mmm:* used to express agreement, to show that the speaker is listening, thinking, like something or is not sure. These were grouped under the headword *mhm*;
18. *yo* used to express disagreement;
19. *uhu* used to express disagreement.

The data used in this analysis were extracted from the Extensible Markup Language (XML) files of the grammatically analysed and tagged version of VOICE

(VOICE POS XML 2.0). The total number of tokens in the corpus in its original form was 1,142,982. From these tokens, I separated the data which were not to be included in the analysis (see Table 4). This was made up of tokens representing pauses, hesitation markers and fillers (i.e. *er* and *erm*), partial words, unintelligible speech, laughter, foreign words, breathing, possessive markers (i.e. *'* and *'s*) and onomatopoeic noises encoded in the International Phonetic Alphabet (e.g. *rrr*, *bəm*, *bʊm*, etc.).

Amongst the possessive markers were seven tokens that had been tagged *POS(POS)/VBS(VBS)*, a tag used to denote that it was not possible for the corpus compilers to disambiguate these (*'s*) items between either the possessive marker or the abbreviated form of *is*. I decided to allocate these seven tokens to the verb *to be* because the number of occurrences of the third person of the verb *to be* in the corpus outnumbers the number of possessive markers by almost fifty to one (i.e. 30,760 to 620). Thus, this allocation had no significant impact on the representation of the respective forms in the corpus. These operations left an overall remaining total number of 934,362 tokens.

Included in the remaining tokens were 2,136 hyphenated compound words. I checked these against the comprehensive Oxford English Dictionary (OED) database online and separated those not found as compounds in the dictionary into their individual constituents and allocated them to their respective word families. Thus, for example, items such as *kick-off*, *eye-catching* and *daughter-in-law* were retained as single tokens because they were present in OED, whilst tokens such as *computer-readable*, *four-metre-high*, and *end-of-the-year* were separated because they were not found in OED as compound words. Additionally, I separated all cardinal and ordinal numbers between twenty-one and ninety-nine. This resulted overall in an additional 1,628 tokens being added to the corpus, bringing the final number of tokens to be included in the analysis to 935,990.

Table 4: Breakdown of VOICE tokens (Version: VOICE POS XML 2.0)

Total initial number of tokens in VOICE	**1,142,982**
Tokens removed before analysis:	
Pauses	112,278
Hesitation markers and fillers	43,527
Partial words	13,395
Unintelligible speech	13,030
Laughter	11,056
Foreign words (non-English speech)	7,906
Breathing	6,602
Possessive markers (i.e ′ and *′s*)	623
Onomatopoeia	203
Tokens remaining	934,362
Tokens added: separated compounds	1,628
Tokens included in the analysis	**935,990**

Next, I examined the group of tokens in the VOICE corpus classified as *Pronunciation variations and coinages* (PVC). This category makes up 0.3% of the total number of tokens retained for analysis (*n*=2,193 tokens), and it is described in the corpus mark-up conventions manual as one which captures "[s]triking variations on the levels of phonology, morphology and lexis as well as 'invented' words" (VOICE Project 2007a: 4). The Oxford Advanced Learner's Dictionary 7th edition (OALD7) was used by the corpus compilers as a reference tool for the compilation of the corpus, and utterances which varied in pronunciation by one or more syllable from entries found in OALD7 were included in the PVC category (Pitzl, Breiteneder and Klimpfinger 2008: 26–27). The corpus compilers note that although these items are "non-codified", they seem "to be communicatively effective" (ibid:22). Moreover, they point out that though some of the items included in this category may be "part of specialized terminology in various disciplines, others appear to be new and innovative" (ibid:22).

In order to identify suitable word families to allocate the PVC tokens to, I began by looking for evidence of current usage of the words. I did this by firstly checking the tokens in this category against the words contained on the BNC/COCA word lists and I found that 28% of the words were present on the lists. Examples of these are *benchmarking*, *bilingualism* and *intercultural*. I then checked the remaining

words against the most comprehensive OED database online, and when not found there, I searched for evidence of them having been used in books published in English by querying Google Books Ngram Viewer. I found 36% of the words referenced in OED as being in current usage (i.e. since the 1970s), for example, *acculturalization*, *annihilator* and *e-learning*. A further 7% of the words were present in books contained in the Google Books database, published in English between 1970 and 2008, for example, *epileptogenesis*, *euroization* and *interrail*. Overall, I found evidence of current usage in the English language for a total of 71% of the PVC tokens (see Table 5), and I allocated these words to suitable word families.

Of the remaining 29% of the PVC tokens, one-fifth were found to be approximations of standard English words, i.e. they were not listed as being in current usage in OED and were not found in Google Books, but they were very similar to standard English usage and could easily be recognised as being slight deviations from standard English words. Examples of these include *anniversity "anniversary", catched "caught"* and *conspirating "conspiring"*. For this group of tokens, I added the word to the word family of its standard equivalent. Of the residual tokens (8% of the PVCs), I categorised 4% as coinages. These were tokens for which I could not find evidence of previous usage, but which seemed to achieve their communicative purpose in the context. Examples of these are *e-education, metacapacity* and *resophistication*. These tokens were also assigned to their respective word families according to the affixation criteria applied to all the data, as described above.

Table 5. Analysis of *Pronunciation variations and coinages* in VOICE

Categories	**Tokens**	**Percentage**
Words found in OED	795	36%
Words found in BNC/COCA lists	621	28%
Words found in Google Ngram	152	7%
Subtotal tokens: evidence found of current usage	**1,568**	**71%**
Approximations of standard English word-usage	455	21%
Possible coinages	87	4%
Unknown words (i.e. none of the above)	83	4%
Subtotal tokens: no evidence found of current usage	**625**	**29%**
Grand total of tokens in PVC category	**2,193**	**100%**

Finally, I assigned the remaining 4% of the PVC data (*n*=83 tokens), of which I was unable to identify the meaning, to a category termed "Unknown". Examples, of such items are (1) *anti-practic* used in the utterance: "some anti-inflammatory usually **anti-practic** agents", (2) *compend* used in the utterance: "are there unique centers for property p are not **compend** well", and (3) *attitunity* used in the utterance "regular growth **attitunity**". These items were retained in the tokens for the overall coverage analysis and were categorised as single constituent "word families".

After the operations described above the data was ready for analysis. The lexical coverage figure was calculated by dividing various frequency levels by the total number of word-families or flemmas in the corpus to obtain the percentage of text coverage. For instance, to arrive at the coverage figure for the most frequently occurring 2,000 word families, I divided the total number of words occurring in those 2,000 families (913,342) by the total number of retained tokens from the corpus (935,990). This resulted in the calculation 913,342/ 935,990= 97.6%.

This methodology was similar to that used in the study by Schonell, Meddleton and Shaw (1956) and replicated by Adolphs and Schmitt (2003). The main differences are in the sizes and types of corpora used and some minor differences in methodology. In terms of corpora, the one compiled by Schonell *et al.* (1956) is approximately half the size of the VOICE corpus, whilst CANCODE corpus, used in the Adolphs and Schmitt (2003) study, is five times larger than VOICE (see Table 6). The language backgrounds of the corpora are ELF for VOICE and native-speaker English for the other two corpora: British and Irish for the CANCODE and Australian for the Schonell, Meddleton and Shaw (1956) corpus. The type of speech is spontaneous interaction in both VOICE and CANCODE, whilst in Schonell *et al.* (1956) around half of the speech is spontaneous interaction and the other half consists of data from interviews between the participants and the researchers. From the point of view of the socioeconomic backgrounds of the participants, CANCODE appears to be the most balanced, representing, according to Adolphs and Schmitt (2003: 427), participants from all segments of society. In contrast, the corpus collected by Schonell *et al.* (1956) represents speech from semi-skilled and unskilled workers. VOICE, instead, appears to be most representative of the higher end of the socioeconomic scale based on my own analysis of the professions of the participants.

Table 6: Comparison of corpora (Schonell *et al.* (1956), CANCODE and VOICE)

	Schonell *et al.* (1956)	**CANCODE**	**VOICE**
Size	Approx. half a million words	Approx. five million words	Approx. one million words
Language background	Australian	British and Irish	Approx. 50 language backgrounds
Type of data	Half spontaneous interaction Half interviews with the researchers	Spontaneous interaction	Spontaneous interaction
Socioeconomic background	Semi-skilled and unskilled workers	All segments of society	Well-educated (own analysis based on participant professions)

In terms of methodology, the word family (i.e. grouping semantically related inflected and derivative word forms under a single headword) was used as a lexical counting unit in all three studies. However, in Adolphs and Schmitt (2003) derivatives formed with prefixes were not included under the same headword, but derivatives formed with suffixes were. Additionally, both in this study (see Table 3) and in Schonell, Meddleton and Shaw (1956) accommodations were made to account for the different meanings of homographs, whilst this was not done in the Adolphs and Schmitt (2003) study. Furthermore, interjections were included in the analysis in both this study and Adolphs and Schmitt (2003), but not in the Schonell, Meddleton and Shaw (1956) study. Moreover, in addition to the word family, both this study and Adolphs and Schmitt (2003) provide figures also for the word type. Finally, a lexical counting unit which is intermediate to the word type and word family (i.e. the flemma) is also used in this study, but was not calculated in either Schonell, Meddleton and Shaw (1956) and Adolphs and Schmitt (2003).

In order to verify the corpus size effects on the results, the lexical coverage was calculated not only for the whole corpus by also using three subsamples: 25%, 50% and 75% of the corpus. The sampling method aimed at maximising the representativeness of the subsample: the corpus files were arranged alphabetically based on the names of the files, which are formed based on the discourse domains (professional, leisure and educational) and speech-event types (interviews, press conferences, service encounters, seminar discussions, working group discussions,

workshop discussions, meetings, panels, question-answer sessions and conversations). For the subsample of 25% of the corpus, every fourth file was included, for the 50% subsample every other file was included and for the 75% subsample every fourth file was excluded. This sampling method guaranteed that all discourse domains and speech-event types were included in the subsample in a way representative of the whole corpus.

3.2.2 Frequency profiling

In addition to calculating the lexical coverage, I frequency profiled the word families occurring in the VOICE corpus against those on the BNC14K word lists and BNC/COCA25K word lists (described in the previous section). The BNC14K word lists "are sequenced largely according to their range and frequency in the 10 million spoken section of the BNC" (Nation 2006: 80). Like the BNC/COCA lists, the words are grouped into word families which include all inflected forms and derivatives formed with affixes up to level 6 on Bauer and Nation's (1993) seven-tier model of morphological affixation (see Table 2 in section 2.2 for details).

The reason for using the BNC lists to frequency profile VOICE was to obtain results comparable to Nation (2006), so that it would be possible to verify how the amount of vocabulary needed to understand English in purely native-speaker contexts compares to those where it is used as a lingua franca. For this same reason, interjections (*n*=18,023) were removed for the frequency profiling analysis, as these were not included in Nation (2006). This brought the total number of tokens to be included in the profiling analysis to 917,967. By also frequency profiling VOICE, I analysed not only what coverage could be achieved by various levels of word family and word type frequency within VOICE, but also how this compares with English language usage amongst native British and American speakers of English represented by the word family frequency data derived from BNC and COCA.

There are only three key differences between this study and Nation (2006): firstly, the range of discourse types included in the data, secondly, the language backgrounds of the speakers and lastly, the size of the corpora. More specifically, the discourse types of the data used in Nation (2006) are of two types: one half of the data is from interviews and talk-back radio (*n*=100,000 tokens), in which listeners phone in with their spontaneous comments on a variety of issues, and the other half

of the data (*n*=100,000 tokens) represents friendly conversation between friends and family. Instead, VOICE corpus covers a much wider range of discourse types.

Secondly, Nation analysed native-speaker data from the Wellington Corpus of Spoken English: the data represented unscripted speech from a range of speakers who had lived in New Zealand since before the age of ten (Holmes, Vine and Johnson 1998). The data in this study is unscripted ELF speech, that is to say the speech of people from around 50 different (mainly European) language backgrounds, including seven per cent of native English speakers. The final difference is that 200,000 tokens were used to calculate the lexical coverage of spoken English in Nation (2006), whilst the VOICE data is five times larger, at around one million tokens.

In order to verify the corpus size effects on the results, the frequency profiling was also calculated not only for the whole corpus by also using three subsamples of one fourth, one half and three fourths of the corpus. The sampling method was the same as that described in the lexical coverage section above.

In addition to using the BNC lists to profile VOICE, I also supplemented the analysis of VOICE by profiling the corpus against the newer and more comprehensive BNC/COCO lists. Furthermore, I analysed what kinds of words occurred at the various levels of frequency. For example, I divided the headwords into two groups: function words and content words. The words which were grouped as function words in this analysis of VOICE were conjunctions, determiners, prepositions, pronouns, auxiliary verbs and *wh*-words. All remaining words were grouped as content words, including adjectives, adverbs, nouns and verbs.

4 Results

In this section, I will first describe the findings for the lexical coverage offered by the most frequent word families, flemmas and word types in VOICE. Then, I will go on to describe the findings for the lexical coverage of VOICE offered by the BNC and BNC/COCA word lists. I will also present my findings for an analysis of VOICE in terms of content versus function words.

4.1 Lexical coverage by word family, flemma and word type

The analysis of lexical coverage in VOICE revealed that to reach 95% lexical coverage in VOICE, 1,204 word families, 1,633 flemmas or 2,598 individual word types are needed (see Figure 1). Instead, to reach 98% coverage, 2,242 word families, 3,259 flemmas and 5,278 word types are required. Thus, approximately twice as many word types, flemmas and word families are needed to reach 98% lexical coverage compared to the number of lexical items needed to achieve 95% lexical coverage.

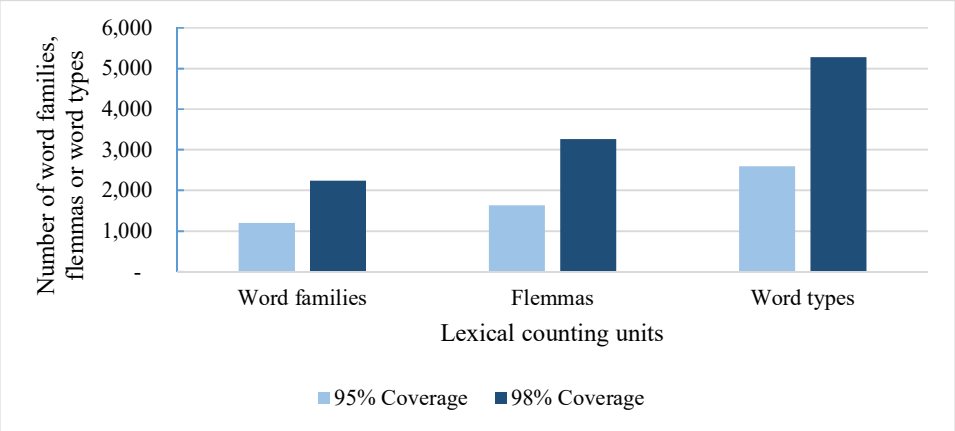

Figure 1: Number of word families, flemmas and word types needed to reach 95% and 98% lexical coverage in VOICE

A much larger proportion of word families, flemmas and word types make up the remaining 2% of the data (see Figure 2). In all, 7,263 word families, 10,396 flemmas and 14,679 individual word types are needed to reach 100% coverage of VOICE. That means that roughly three times as many words families, flemmas and word types are needed to reach 100% lexical coverage than the number needed to reach 98% lexical coverage.

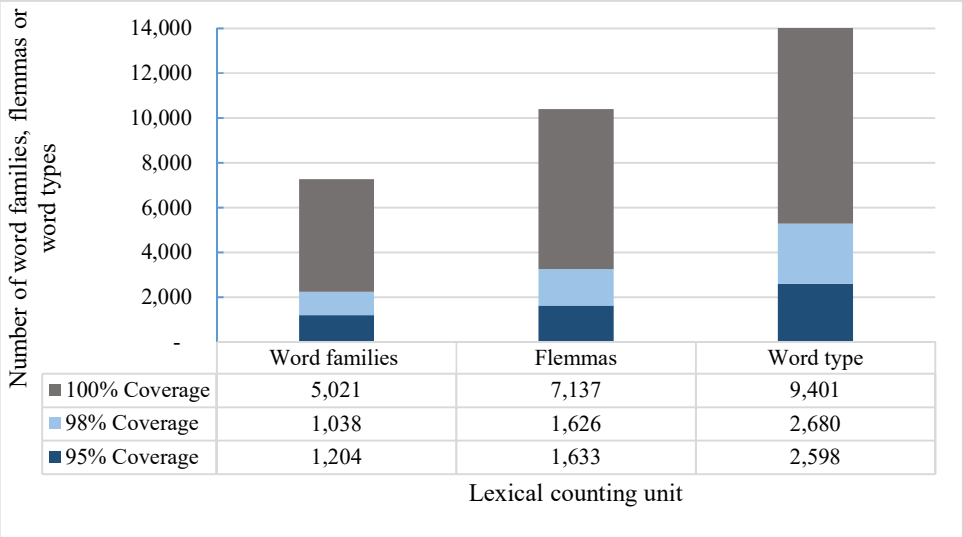

Figure 2: Overall number of word families, flemmas and word types in VOICE (total n=7,263 word families, n=10,396 flemmas and n=14,679 word types)

4.1.1 Corpus size affect on lexical coverage

In order to ascertain whether and how the corpus size might affect the lexical coverage results, I checked the figures for varying representative subsamples of VOICE: 25% (*n*=240,006 tokens), 50% (*n*=473,539 tokens), 75% and 100% (*n*=727,496 tokens) of the corpus (see Figure 3). This analysis shows that the sample size has only a negligible effect below a certain size (around 700,000 tokens), but above that size it does not appear to have any significant impact on the coverage figures.

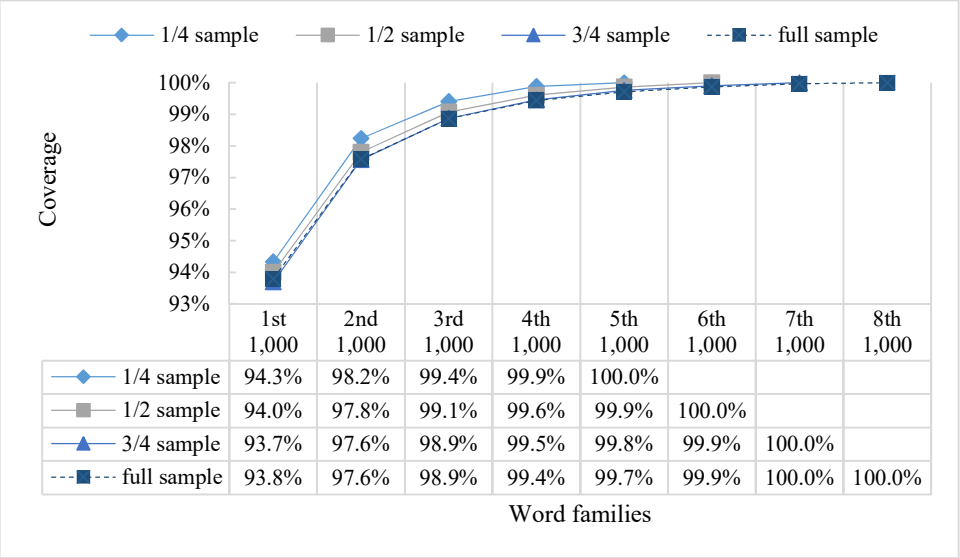

Figure 3: Lexical coverage in VOICE for varying sizes of subsample compared to the whole sample (n=935,990 tokens)

4.2 Frequency profiling of VOICE against BNC and BNC/COCA word lists

The total number of tokens used in the frequency profiling was 917,967, i.e. all of the tokens included in the calculation of lexical coverage (*n*=935,990), minus interjections (*n*=18,023). The frequency profiling of VOICE against the BNC lists (see Figure 4 and Table 7) revealed that the BNC's first 1,000-word family list offered 90.9% lexical coverage of VOICE (excluding proper nouns) and 92.2% (including proper nouns). The frequency profiling against the BNC/COCA lists showed that the BNC/COCA's first 1,000-word family list offered 2.3% less coverage than the BNC's first 1,000-word family list at 87.5% excluding proper nouns and 89.9% including proper nouns.

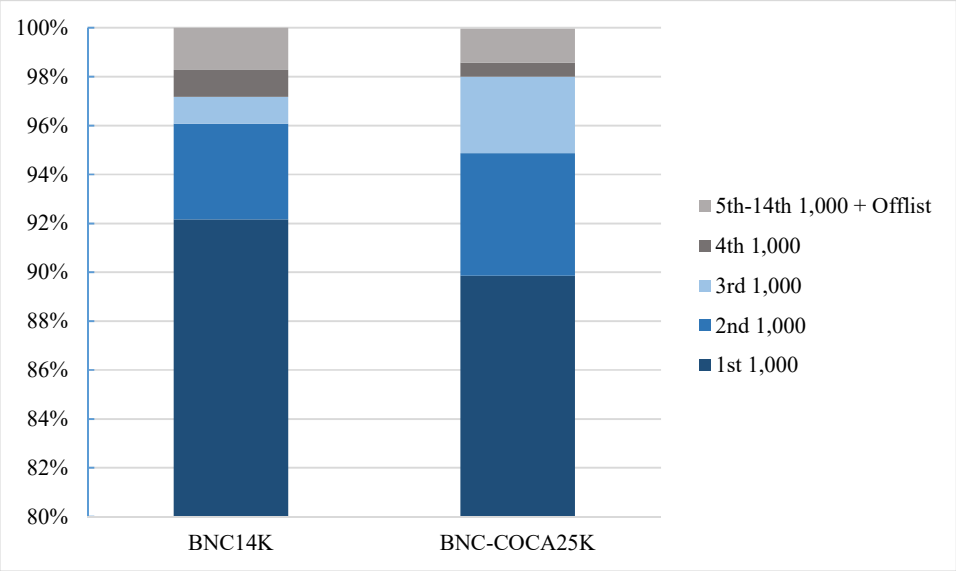

Figure 4: Frequency profile of VOICE against BNC and BNC/COCA word family lists (total *n*=917,967 tokens)

With the second 1,000 most frequent BNC word families, a further 3.9% coverage was gained, bringing the overall coverage at this level to 96.1% (including proper nouns). The same level of the BNC/COCA lists offered 5% coverage, for an overall coverage of 94.9% (including proper nouns). Thus, the two thousand most frequent word families of the BNC/COCA word lists offered 1.2% less coverage than the two thousand most frequent word families of the BNC lists.

With the third 1,000 most frequent BNC word families, lexical coverage of 97.2% of VOICE was reached, whilst 98% coverage was reached with the BNC/COCA lists at this level. Hence, the coverage of the BNC/COCA lists at the

3,000 word family level surpassed that offered at the same level of the BNC lists by 0.8%.

The forth 1,000 most frequent word families of the BNC offered a further 1.1% lexical coverage of VOICE, for an overall coverage at this level of 98.3%, whilst only 0.6% additional coverage was gained with the BNC/COCA lists at this level, for a total coverage of 98.6% including proper nouns. The 5th to 14th BNC levels together offered 1.5% lexical coverage of VOICE, and the remaining 0.3% of tokens (*n*=2,320) were made up of word families (*n*=500 accounting for 2,320 tokens) not resulting on the BNC word family lists. The 5th level upwards of the BNC/COCA lists offered 1.3% coverage, and 0.1% (*n*=504 tokens and 107 families) were words that did not appear in VOICE.

Table 7: Profiling of VOICE corpus against the BNC and BNC/COCA word family lists (total *n*=917,967 tokens)

	BNC word lists			BNC/COCA word lists		
Word lists	**Tokens**	**Word types**	**Word families**	**Tokens**	**Word types**	**Word families**
1st 1,000	90.9%	4,376	991	87.5%	3,593	999
2nd 1,000	3.9%	2,894	963	5.0%	2,743	926
3rd 1,000	1.1%	1,574	766	3.1%	2,594	927
4th 1,000	1.1%	1,225	588	0.6%	1,122	643
5th 1,000	0.5%	856	485	0.3%	645	444
6th 1,000	0.3%	512	362	0.2%	489	328
7th 1,000	0.2%	406	261	0.1%	304	233
8th 1,000 onwards	0.5%	1,154	862	0.7%	1,597	1,405
Proper nouns	1.3%	1,102	1,079	2.4%	1,445	1,235
Not on the lists	0.3%	563	500	0.1%	114	107
Total	**100%**	**14,662**	**6,857**	**100%**	**14,646**	**7,247**

In terms of the number of tokens, word types and word families occurring in VOICE at each level of frequency on the BNC lists, the frequency profiling showed that the words in VOICE are spread over the 14 BNC frequency levels and beyond

(see Table 7). The first 1,000 word families of the BNC lists account for the overwhelming majority (*n*=834,113, i.e. 90.9%) of the tokens in VOICE. These tokens are made up of 4,376 individual word types, which are grouped under 991 headwords on the BNC lists.

Thus, only 9 word families in the first 1,000 word families of the BNC lists were not present in VOICE. These were mostly typically British words, such as *bloke*, *chap*, *lad*, *pence*, *quid* and *wee*. These culturally specific word families, which were on the BNC lists but absent in VOICE, reflect the cultural bias of the reference corpus, the BNC, as well as revealing a weakness of using the BNC lists to make comparisons between ELF and NSE more generally. In addition to these word families, also the word family *rail,* and the letters *v* and *w* from the BNC's 1st 1,000-word list, did not occur in VOICE.

Instead, only one of the word families from the 1st 1,000 word list of the more inclusive and updated BNC/COCA word lists was absent from VOICE, i.e. the word family *engine*. In fact, the nine word families from the BNC's 1st 1,000 word list that were absent from VOICE have been recategorised to lower frequency lists in the BNC/COCA word lists: all letters have been listed as "marginal words", *quid* has been moved to the 8th 1,000-word list, *bloke* to the 7th, *chap*, *lad* and *wee* to the 5th, *rail* to the 3rd and *pence* has been placed under the headword *penny* and remains amongst the 1st 1,000 most frequent word families also on the BNC/COCA lists.

The 3.4% less coverage offered by the BNC/COCA's 1st 1,000 most frequent word families compared to those of the BNC (i.e. 87.5% coverage of VOICE with BNC/COCA's 1st 1,000 word list compared to 90.9% coverage with the 1st 1,000 word list of the BNC) is partly explained by the fact that many words which are categorised in the BNC 1st 1,000 word family list were recategorised as proper nouns in the BNC/COCA lists, e.g. all the names of countries and their associated adjectives. Many of these had a high level of occurrence in VOICE: for example, the word family grouped under the headword *Europe* (which includes the word types *Europe*, *European* and *Europeans*) ranks 86th and accounts for 1,516 tokens in VOICE, whilst the word family grouped under the headword *English* (which includes the word types *England*, *English*, *Englishes* and *Englishman*) ranks 89th and accounts for 1,455 tokens in VOICE. Indeed, it is largely for this reason that overall proper nouns reached 1.1% more coverage of VOICE with the BNC/COCA word lists than with the BNC word lists (i.e. 2.4% and 1.3% respectively).

In addition to this, many other words from the BNC word lists which have a high frequency of occurrence in VOICE were recategorised to lower level frequency lists in the BNC/COCA word lists: for example, the word families *language* (rank = 84th and frequency = 1,543) and *university* (rank = 94th and frequency = 1,335) were moved from the 1st 1,000 word families on the BNC lists to the 2nd 1,000-word list on the BNC/COCA word lists. Indeed, of the 209 word families from the 1st 1,000 word families of the BNC lists that were recategorised to lower levels in the BNC/COCA lists, almost half rank above 750 and occur more than 100 times each in VOICE.

The second 1,000 word families of the BNC lists account for a further 3.9% of the tokens (*n*=35,930) in VOICE. The source of these tokens are 963 word families and 2,894 word types. The 46,135 tokens present in VOICE at the same level of the BNC/COCA lists are made up of 2,743 types and 926 families. More than half of the 37 words families from the second 1,000 BNC list that do not appear in VOICE have been recategorised in the BNC/COCA word lists. The following is a list of these words with the recategorisation (where applicable) in the BNC/COCA word lists shown in brackets:

> *bin, bloom, chuck* (5), *cough, diagram* (4), *drag, drawer, flatting* (5), *garage, inch, jack* (proper noun list and the inflected forms: *jacked*, *jacking* and *jacks* on 8th 1,000-word list), *landlord* (4), *lorry* (8), *midland* (proper noun), *miner* (3), *muck* (7), *muscle, nick* (8), *nil* (8), *nought* (4), *op, parish* (3), *pat, pint* (5), *pit* (3), *pudding, pump, redundant* (5), *repair, rob, sack* (4), *sandwich, sod* (marginal word), *sub* (removed from BNC/COCA), *tidy* (5), *tory* (proper noun) and *ward* (4).

Several of these words also have a British (or American) cultural or regional bias, for example, *inch*, *lorry*, *Midland*, *nil, nought*, *pint*, *pudding, sod* and *Tory*, so it is not surprising that they did not occur in an ELF corpus such as VOICE. On the other hand, the absence of other words may be a little more surprising, such as *cough, muscle, repair* and *sandwich*.

The 74 word families from the BNC/COCA's second thousand most frequent word families that were absent in VOICE were the following:

ace, angel, anger, bacon, bark, bin, blanket, bleed, bloom, bow, bucket, bump, cage, canoe, cape, captain, casual, centimetre, cheek, cop, cotton, cough, crawl, creature, creep, dawn, dine, drag, drawer, fox, frog, fur, garage, grin, heap, inch, jaw, knit, lamb, lamp, lawn, leap, lid, mow, muscle, nest, oak, pat, pine, pudding, pump, repair, spray, rice, roar, rob, sandwich, shade, shiver, snake, snap, sorted, steak, steam, stiff, storm, thief, towel, trunk, wander, weed, wicked, wolf and *wool.*

It may be of interest to note that around 20 per cent of the words that are ranked on the BNC/COCA's second thousand words list but are absent from VOICE appear to be related to nature: *bloom, creature, fox, frog, lamb, lawn, mow, nest, oak, pine, snake, weed* and *wolf*.

The 1.1% coverage of VOICE offered by the third 1,000 word families[1] of the BNC lists is accounted for by 10,073 tokens (i.e. three and a half times fewer tokens than at the 2nd 1,000 word level). The source of these tokens are 1,574 word types and 766 word families. Instead, 927 word families and 2,594 word types account for 28,715 tokens (i.e. 3.1%) at the same level of the BNC/COCA list.

The 232 word families from the BNC's 3rd 1,000 word family list that were absent in VOICE are the following:

abbey, aerial, affection, alley, almighty, aluminium, anger, appal, arrears, avenue, badge, bark, barn, bash, beam, bench, blanket, ***blimey****, bog, bolt, boo, borough, bow, brass, bucket, bully, bump, burgle, canal, candle, captain, casual, casualty, cathedral, cement, chapel, cheek,* ***cheerio****, clutch, collar, congregate, cop, cord, cotton, cracker, cramp, crawl, creep, cricket, cripple, crush, crystal, daft, damp, derby, detach, dine, dinosaur, disgrace, distress, ditch,* ***dodgy****, doorstep,* ***draught****, dread, dreadful, drill, drip, dye, eldest, escort, fatal, felled, fiddle, fir, flap, flare, ford, fume, funeral, fur, furnish,* ***gallon****, glaze, glow, gospel, grief, gut, hassle, hay, heal, heather, helicopter, hen, hip, hockey, hood, hooray, horrendous, humour, idle, incline, indulge, inn, jaw, jewel, jolly, jug, kettle, knot, lamb, lamp, lawn, leaf, leap, lid, litter, lodge, loft,* ***loo****, manor, mar, merchant, mild, misery, moan, mock, motorbike, mug, nest,*

[1] There are, in fact, only 998 word families on the 3rd 1,000 BNC word list.

nip, nowt, nuisance, oak, obscure, outrage, overtake, overtime, paddy, pale, palm, par, pathetic, pigeon, pinch, plaster, ***plonk****, plough, poke, pond, potter, preach, princess, query, rattle, receipt, rescue, resort, ribbon, rotten, rugby, scrap, scrape, scribble, scrub, sergeant, shade, shovel, silk, sincere,* ***sixpence****, slap, slim, smack, snap, sniff, spark, spectacle, spit, splash, spray, squash, stab, steam, stiff, storm, strap, stride, suite, supper, surgeon, suspicious, swan, tack, temper, terrific, thief, thrill, thumb, thunder, tilled, timber, token, torch, towel, tragedy, tread, tumble, tyre, undo, upwards, utility, vacuum, vandal, vat, vet, wagon, wander, warrant, weed, whack, whereabouts, whereby, whoop, wicked, widow, wig, wolf, wool, wreck, wrestle* and *wrist.*

Some of these words are also typical of British English (see the bolded words).

The following 73 word families from the BNC/COCA's 3rd-1,000 word family list were absent in VOICE:

acre, adolescent, affection, allege, missile, anxiety, assault, atom, bacterium, beam, bench, blast, blend, circuit, companion, condemn, crush, crystal, damp, defendant, discreet, dna, doctrine, drill, embrace, endure, episode, fabric, firms, flesh, funeral, glow, gravity, halt, hazard, heal, highway, hip, humour, invasion, jail, laughter, leather, lodge, loyal, marine, mild, miner, outrage, pale, palm, parish, pit, psychiatry, raid, rail, render, rescue, resort, seize, senses, shrug, sigh, silk, studio, supreme, swell, thrill, tragedy, utility, veteran, weave and *whisper*.

At the 4th 1,000 level of the BNC word lists, 10,040 tokens (i.e. 1.1%) are present in VOICE. These are made up of 1,225 word types and 588 word families. At the same level of BNC/COCA word lists, the figures are 5,351 tokens (i.e. 0.6%), 1,122 word types and 643 word families.

At the remaining 5th to 14th 1,000 levels of the BNC word lists, 13,558 tokens (i.e. 1.5%) occur in VOICE. The source of these tokens are 2,928 word types and 1,970 word families. The remaining 5th to 25th 1,000 word levels of the BNC/COCA word lists plus the additional word lists (i.e. marginal words, transparent compounds

and abbreviations) account for 12,354 tokens. These tokens are made up of 3,035 word types and 2,410 word families.

Proper nouns account for 11,933 tokens, 1,102 word types and 1,079 word families with the BNC word lists and 21,709 tokens, 1,445 word types and 1,235 word families with the BNC/COCA word lists. As discussed above, the difference in the number of proper nouns on the two lists is mostly because names of cities and countries were placed on the frequency grouping lists in the BNC word lists, whilst they were categorised as proper nouns on the BNC/COCA word lists.

A number of items occurred in VOICE that were not on the BNC (*n*=563 word types and *n*=500 word families) and BNC/COCA (*n*=114 word types and *n*=107 word families) word lists. Examples of some of the most frequent word families that occurred in the VOICE corpus but not on the BNC and BNC/COCA word lists are the following (the number of tokens is shown in brackets and those also not appearing on the BNC/COCA lists are shown in bold):

> ***lingua franca*** (243), *email* (144), *internet* (102), *ngo* (97), *website* (87), *pr* (52), *brainstorm* (23), ***pharmacoresistant*** (23), *hippocampal* (22), ***quaternionic*** (17), *pesto* (16), *proofread* (16), ***epileptogenesis*** (14), ***neuropa*** (14), *conformal* (13), ***cytokine*** (13), *nacho* (13), *sim* (13), *ciao* (12), *fora* (12), *mri* (12), *pretzel* (12), *rapporteur* (12), *snowboard* (12), *epsilon* (11), *melange* (11), *acculturate* (10), ***bomboclat*** (9), ***plurisubharmonic*** (9), webmail (9), ***euroization*** (8), *landline* (8), *mountainboard* (8), *isomorph* (7), ***amygdala*** (6), *feta* (6), *interrail* (6), *menthol* (6), *sensitize* (6), *raki* (6), ***epileptogenic*** (5), *habilitate* (5), *ip* (5), ***metaevaluation*** (5), ***neuroprotection*** (5), *pos* (5), ***poutine*** (5), *quadripartite* (5), *teleconference* (5), *tilde* (5), *webcam* (5)

4.2.1 Corpus size affect on frequency profiling of VOICE

Since the frequency profiling analysis was a partial replication of Nation (2006), it was necessary to verify whether the results found in this study would change if a smaller corpus sample were analysed. Therefore, I took a subsample of the VOICE corpus of a size (*n*=197,422) comparable to the corpus analysed by Nation (2006). In compiling the subsample of VOICE, I also took into account the narrower selection of discourse types included in Nation (2006) compared to those included in VOICE:

thus, I included only question and answer sessions, interviews and conversations in the speech-event types of the subsample, as these seemed the most comparable to the talk-back radio and conversation data included in Nation's (2006) 200,000-word subsample of the Wellington Corpus of Spoken English.

The results of this analysis indicated that the different corpus size and narrower discourse types had, if any, only a negligible effect (see Figure 5) on the lexical coverage of VOICE offered by the BNC lists. Compared to the whole corpus, the results for the subsample showed an increase of only 0.2% in the coverage of the 2nd 1,000 and 3rd 1,000 word families (plus proper nouns), whilst the 4th 1,000 word families (plus proper nouns) reached the same level of coverage and the 6th 1,000 (plus proper nouns) was 0.1% lower.

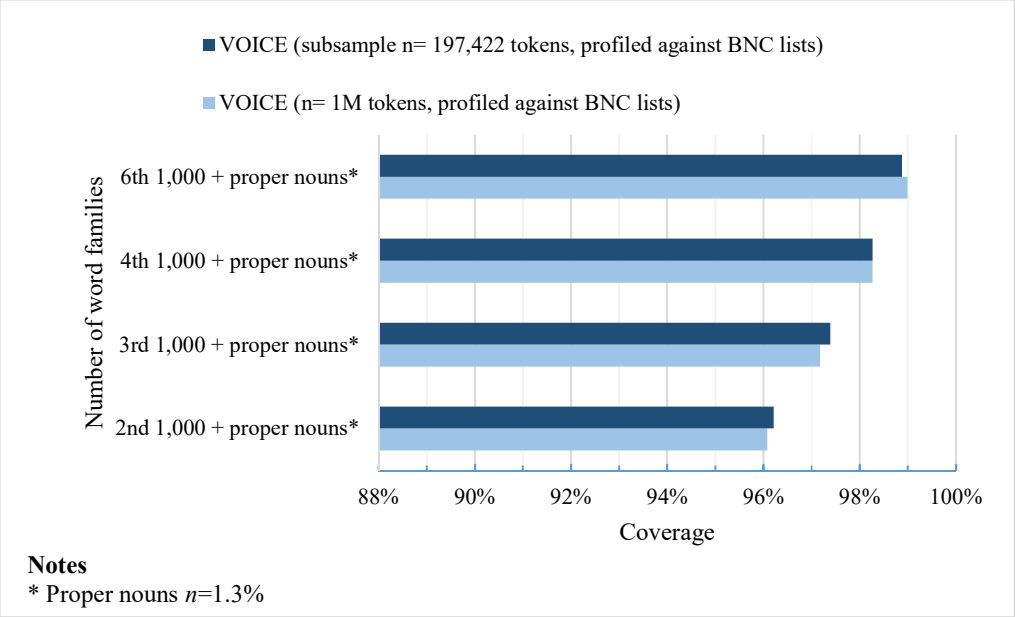

Figure 5: Frequency profiling of VOICE (full sample and subsample) against BNC lists

4.3 Analysis of function words *versus* content words in VOICE

The analysis of the words in terms of content versus function words (see Table 8) revealed that less than one percent of the individual word types in VOICE (i.e. *n*=113 out of a total of *n*=14,646) accounted for almost half (46%) of all the tokens in the corpus.

Table 8: Coverage of content and function word types in VOICE

Category	Frequency grouping	Tokens	Word types	Coverage
Function words	1st 1,000	421,788	100	45.95%
	2nd 1,000	700	6	0.08%
	3rd, 4th, 5th & 6th 1,000	153	7	0.02%
Content words	1st 1,000	381,411	3,493	41.55%
	2nd 1,000	45,435	2,737	4.95%
	3rd 1,000	28,576	2,590	3.11%
	4th 1,000	5,350	1,121	0.58%
	5th 1,000 onwards	12,341	3,033	1.34%
	Proper nouns	21,709	1,445	2.36%
	Not on the lists	504	114	0.05%
Total		**917,967**	**14,646**	**100.00%**

Most of the function words (n=100 word types and n=421,788 tokens) occurred amongst the first 1,000 most frequent word families, whilst the remaining 13 function word types, occurring amongst the 2nd to 6th 1,000 word families, accounted for only 853 tokens or 1% of all the tokens in VOICE (see Table 8 for details).

5 Discussion

5.1 Lexical coverage

The current research consensus is that as much as 6,000–7,000 of the most frequent word families may be needed to understand spoken English (Nation 2006, see also Schmitt *et al.* 2017 for a review). These findings are based on studies into the language of native speakers of English (see in particular Adolphs and Schmitt 2003, Nation 2006). The findings of this study suggest that substantially less vocabulary may suffice to understand English in international ELF contexts compared to what has been found for intranational native-speaker contexts. The present study, which uses ELF data, is a partial replication of studies that have used data from intranational native-speaker contexts. The first of these studies (Adolphs and Schmitt 2003) was itself a replication of a much earlier study (Schonell, Meddleton and Shaw 1956). The lexical coverage figures found for the most frequent 2,000 word families varies in the three studies (see Figure 6): 99% in Schonell, Meddleton and Shaw (1956), 98% in the present study and 95% in Adolphs and Schmitt (2003).

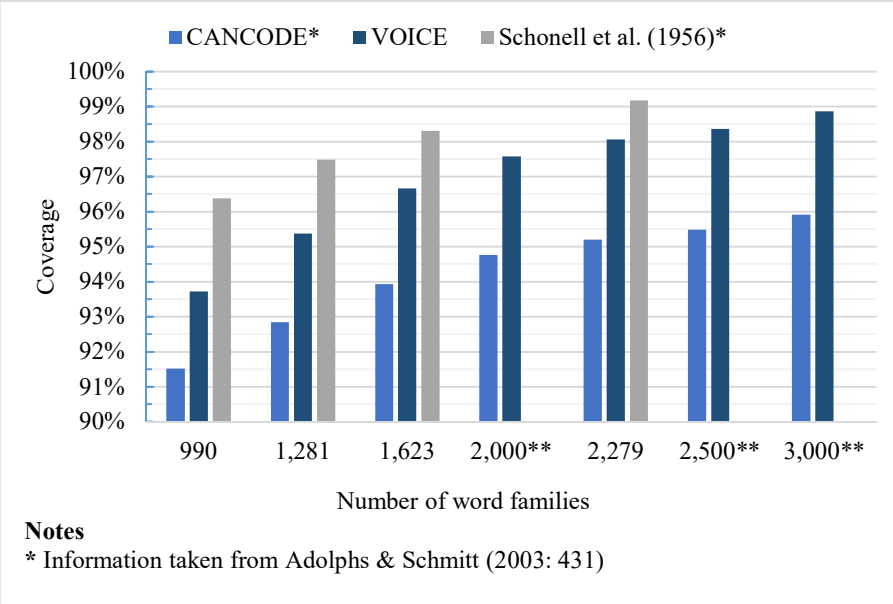

Figure 6: Lexical coverage in VOICE compared to Schonell *et al.* (1956) and CANCODE

It is clear from Figure 6 that the trend is consistent for the three studies at each threshold measured, from 1,000 to 3,000 word families: CANCODE offers the least coverage (4–5% less than the Schonell *et al.* (1956) corpus and 2–3% less than VOICE). VOICE offers greater lexical coverage than CANCODE but less than Schonell *et al.* (1956) (approximately 1–3% less), and the Schonell *et al.* (1956) corpus offers the greatest lexical coverage of the three corpora.

The discrepancies in lexical coverage figures found for the three studies are explained only in minimal part by the different corpus sizes: CANCODE has five million tokens, VOICE has one million tokens and the Schonell *et al.* (1956) corpus had half a million tokens. In fact, the representative subsample of VOICE of a size comparable to Schonell *et al.* (1956) (*n*=473,539 tokens) provided higher coverage of only 0.2% at all frequency levels from the 1st 1,000 words to the 5th 1,000 words (see Figure 3 in section 4.1.1). This indicates that the difference in size would explain at most 0.2% of the 1.1 % higher lexical coverage found for the Schonell *et al.* (1956) corpus compared to the VOICE corpus: i.e. 2,279 word families offered 99.2% lexical coverage with the Schonell *et al.* (1956) corpus versus 98.1% lexical coverage with VOICE.

This finding is confirmed by Adolphs and Schmitt (2003: 435–436), who also found no substantial difference in coverage figures when they checked the lexical coverage of a representative subsample of the CANCODE corpus which was of a

comparable size (497,658 tokens) to the Schonell *et al.* (1956) corpus. Additionally, since the discrepancy in lexical coverage found for the subsamples of VOICE disappears above 700,000 tokens (see Figure 5 in section 4.2.1), it suggests that the corpus size difference between VOICE (one million words) and CANCODE (five million words) does not explain the differences in lexical coverage found for these two corpora. For example, the most frequent 2,000 words in VOICE provided almost 3% higher lexical coverage compared to what was found for CANCODE at the same level, i.e. 97.6% versus 94.8% respectively.

Hence, if the varying corpus sizes explain only a negligible part of the differences in lexical coverage found for the three corpora, then it seems that the principal explanatory factor is the type of data being studied. The Schonell *et al.* (1956) corpus was made up of the spoken interactions of Australian semi-skilled and unskilled workers (n=500,000 tokens). The lexical coverage figures found for the Schonell *et al.* (1956) corpus shows that this group of people used a more limited range of vocabulary in their interactions compared to both the ELF speakers sampled in VOICE and the native English speakers represented by CANCODE.

CANCODE is arguably a much more representative sample of general spoken native-speaker English than *Schonell et al. (1956)*: it is a five-million-token corpus of spontaneous, spoken interactions in a wide variety of discourse contexts and speech genres collected from diverse settings across the UK and Ireland between 1994 and 1999. Thus, Adolphs and Schmitt (2003: 430–432) argue that their findings "are likely to be more representative of the kind of spoken discourse the typical native speaker or L2 learner would be in contact with, simply because CANCODE corpus is a larger, more modern and more diverse sample of general spoken English." I would argue, instead, that because of the current status of English as a global lingua franca, the findings of the present study, using the VOICE corpus, better reflect the kind of everyday, spoken discourse that L2 learners of English are likely to encounter most often. Indeed, VOICE provides the largest currently available sample of general English spoken as a lingua franca in Europe.

It is, however, questionable whether VOICE provides a truly representative and generalisable sample of ELF. For one thing, the corpus size of one million tokens, though the best currently available sample of general ELF, is rather small by today's standards for general corpora. For example, COCA, which provides the largest currently available sample of contemporary American English, currently stands at

560 million tokens, of which 20% (*n*=118 million words) are transcripts of unscripted conversation. On the other hand, the spoken data in COCA is collected in a narrow range of settings (i.e. TV and radio) compared to VOICE's broader range of settings (i.e. professional, educational and leisure), which also calls into question whether COCA's spoken section can truly be considered a generally representative sample of contemporary, spoken American English.

A second concern with how representative VOICE is of general ELF is that upon examination of the data it appears to have an academic bias. For example, the following words related to academia were amongst the top 1,000 most frequent word families in VOICE, whilst they are at decidedly lower frequency level on the BNC lists (the relevant BNC list is indicated in brackets): academy (4th 1,000), professor (4th 1,000), tutor (4th 1,000), thesis (6th 1,000), bachelor (7th1,000), rector (9th 1,000), semester (9th 1,000) and PhD (14th 1,000). This academic bias in VOICE is also confirmed by the distribution of the participants' occupations: four out of ten participants are students, and 7% of the participants are university employees, including professors and lecturers. Additionally, the professions of all but a small minority (1%) of the participants were skilled or highly-skilled, suggesting that VOICE is more representative of a highly-educated section of society rather that society more generally.

Based on the results of this study, VOICE participants used a wider range of vocabulary than the Australian workers in the Schonell *et al.* (1956) corpus. This is probably explained by the high level of education of the VOICE participants compared to the blue-collar workers in the Schonell *et al.* (1956) corpus. Yet, though VOICE appears to be more representative of a highly-educated population of ELF speakers, their range of vocabulary is markedly lower than the general native-speaker population represented by the sample in CANCODE. These results indicate that a much smaller number of word families are needed to understand English in international contexts where it is spoken as a lingua franca compared to intranational contexts where it is used amongst native speakers of English.

5.2 Frequency profiling

The method for analysing the lexical coverage of VOICE discussed in the previous section only takes into account the words in the corpus itself. The second method of analysis was aimed at verifying how this compares to English usage more generally

by comparing the frequency ranking of word families found in VOICE to that of 14 1,000 word family lists ranked according to frequency, range and dispersion data from on the BNC. Additionally, since these lists have since been supplemented with frequency, range and dispersion data from COCA resulting in twenty-five 1,000 ranked word family lists, I complemented this analysis by profiling VOICE against these word family lists too.

As such, this study is a partial replication and extention of Nation (2006), who first carried out such an analysis when he profiled a 200,000-word subsample of the Wellington Corpus of Spoken English (WCSE) against the BNC14K word family lists. Based on his findings the general consensus amongst scholars has been that between 6,000–7,000 word families (plus proper nouns) are needed to understand spoken English. However, Schmitt *et al.* (2017) have called for the need to replicate and validate these findings.

The results of these methods of analysis with VOICE (see Figure 7) revealed that a much higher lexical coverage is achieved with far fewer word families in VOICE when profiled against the BNC word lists than Nation (2006) found for the Wellington Corpus of Spoken English: around 4,000 of the most frequent word families (plus proper nouns) make up 98% of all the word families in VOICE. With the more modern and comprehensive BNC/COCA word family lists even fewer word families are needed to achieve 98% lexical coverage: just 3,000 of the most frequent word families (plus proper nouns).

Using frequency-ranked word lists to determine how much vocabulary is needed to understand spoken English is based on the assumption that people tend to learn more frequent words before less frequent ones due to the likelihood of greater exposure to them (see, for example, Read 1988 and Laufer *et al.* 2004).

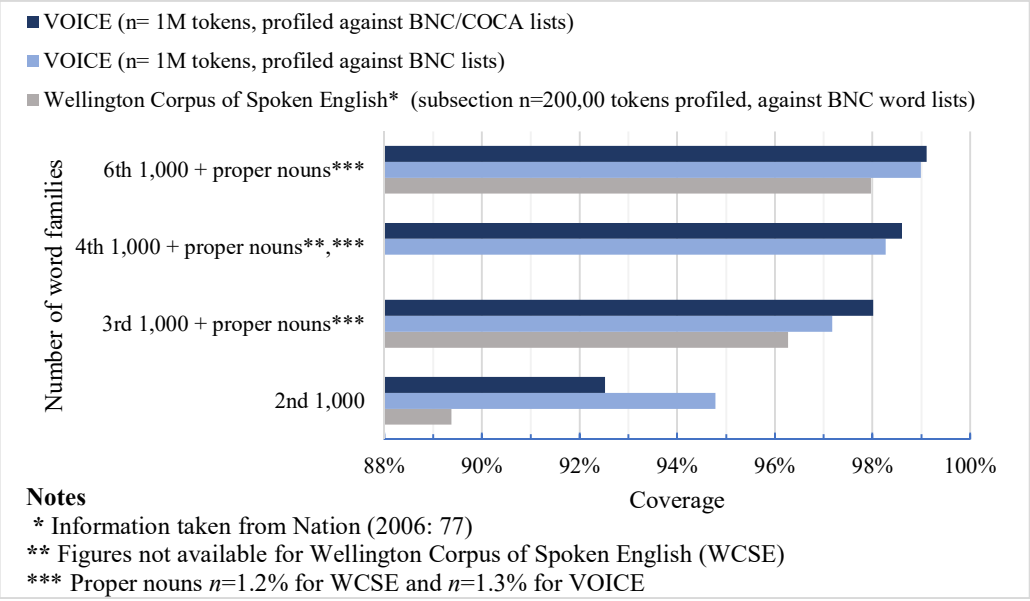

Figure 7: Frequency profiling of VOICE against the BNC and BNC/COCA lists compared to Nation (2006)

The better coverage offered by the BNC versus BNC/COCA lists at the 2,000-word level is accounted for by the fact that 238 of the word families (*n*=21,639 tokens) occurring amongst the most frequent 2,000 word families in VOICE and on the BNC 1st and 2nd 1,000-word lists were recategorised to the BNC/COCA 3rd–9th 1,000 word lists, as well as to the additional lists (proper nouns, marginal words and transparent compounds). This discrepancy is in small part offset by 45 word families (*n*=1,997 tokens) that occurred amongst the top 2,000 word families in VOICE and that were on the lower frequency BNC lists (i.e. 3rd–4th 1,000 word lists), but were recategorised to the 1st and 2nd 1,000-frequency level on the BNC/COCA word lists. Additionally, two new word families that were not present in the BNC lists, *email* and *internet* (accounting for *n*=246 tokens in VOICE), were added to the more recent BNC/COCA lists.

Furthermore, 150 word families (*n*=9,868 tokens) in VOICE representing the names of countries and religions were recategorised as proper nouns in the BNC/COCA lists. This is the reason why proper nouns make up 1.1% more tokens when VOICE is profiled against the BNC/COCA lists (2.4%) than when the corpus is profiled against the BNC lists (1.3%).

The differences in coverage offered by the first three thousand most frequent words of the BNC/COCA lists compared to the BNC lists are also the result of recategorisations of words between the two lists. Indeed, 218 word families

(accounting for *n*=7,882 tokens in VOICE) that occurred amongst the top 3,000 word families in VOICE and are present on the 1st to 3rd BNC/COCA word lists had been recategorised from lower frequency levels (i.e. 4th to 9th) on the BNC word lists. The analysis of VOICE would appear to confirm this recategorisation.

Apart from these differences, the verification of the effect of the sample size and discourse types (see Figure 8) indicates that these factors explain at most a negligible amount of the greater lexical coverage offered by the BNC list profiling of VOICE compared to the BNC list profiling of the subsample of the Wellington Corpus of Spoken English (Nation 2006). Hence, this suggests that the main reason for the differences found is related to the spoken discourse of native speakers versus English spoken as a lingua franca in international contexts. This confirms the trend of the results of the first analysis of the lexical coverage of VOICE, i.e. that far fewer word families are needed to understand English in contexts where it is spoken as a lingua franca.

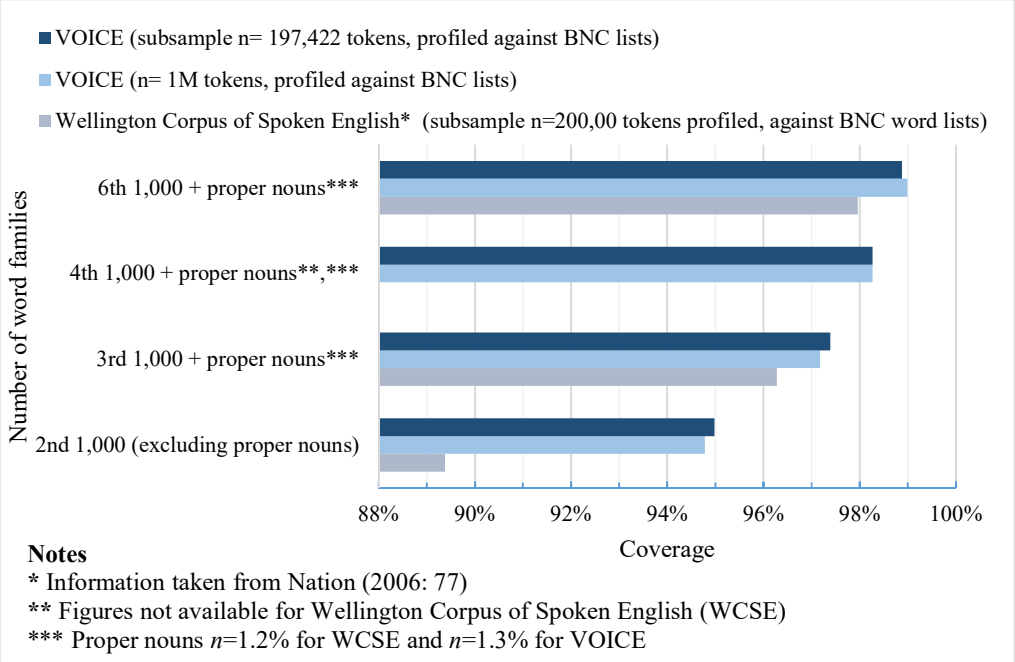

Figure 8: Frequency profiling of VOICE (full sample and subsample) against BNC lists compared to Nation (2006)

However, with the first analysis of VOICE, when only the frequency of the word families within the actual corpus was considered, around 1,000 word families (plus proper nouns) were required to reach 95% and 2,000 word families (plus proper nouns) offered 98% coverage. Instead, when VOICE is profiled against the BNC and

BNC/COCA lists roughly twice as many word families are needed to reach the same levels of coverage: i.e. 1,000–2,000 for 95% coverage and 3,000–4,000 word families (plus proper nouns) for 98% coverage. Though the latter methodology used to analyse VOICE results in twice as many word families needed to gain the same levels of coverage compared to the former methodology, this is still approximately half the number of word families that Nation (2006) found for the native speakers of English in the Wellington Corpus of Spoken English: i.e. Nation (2006) found that 2,000–3,000 word families (plus proper nouns) offer 95% lexical coverage and 6,000–7,000 word families (plus proper nouns) are needed to reach 98% lexical coverage. This implies that L2 learners of English whose aim it is to understand spoken English in international settings where it is used as a lingua franca will need half the amount of vocabulary they would need to understand spoken English in native-speaker, intranational contexts. This is a remarkable saving for L2 learners of English, especially considering that L2 learners of English appear to often fail to reach the levels of vocabulary knowledge which Nation (2006) claims is needed to understand spoken English (see Schmitt 2008: 332 for a review).

5.3 Level of lexical coverage required for listening comprehension

The question remains, however, whether it is 95% or 98% lexical coverage that would provide learners of English with adequate lexical resources to understand spoken English in a wide variety of settings. At 95% lexical coverage, the listener would have to deal with between five unknown words in every hundred or around 6–8 words per minute, at an average speaking speed of 110–150 words per minute. Instead, 98% lexical coverage means that listeners would face two unknown words in every 100 or three unknown words per minute of speech assuming an average speaking speed of 150 words per minute.

Obviously, the more vocabulary a learner knows the better, but the question is: what is the minimum threshold of lexical knowledge needed to understand a wide variety of spoken discourse? Or put differently, how many unknown words can be tolerated before comprehension breaks down? This, of course, depends of the specific demands of the context, with some situations likely to require a higher command of vocabulary knowledge than others. For example, in a university lecture, where the flow of information is mostly unidirectional, it is likely that the ability to quickly and efficiently decode lexical information will be needed to secure

comprehension. Similarly, when listening to a radio news broadcast, the speed of speech and the lack of non-verbal cues is also likely to put a premium on higher lexical coverage and the quick online processing of lexical information than would be required for more interactive situations. For example, in a face-to-face conversation between friends, gestures and facial expression can facilitate comprehension. Additionally, meaning can more easily be negotiated through clarification, rephrasing and confirming understanding. Thus, it is probable that in such contexts, where compensatory strategies can be deployed, more unknown vocabulary can be tolerated without impeding comprehension.

Research that has investigated the interaction between lexical coverage and the ability of L2 learners of English to understand spoken discourse has found a minimal threshold of 95% and an optimal threshold of 98% depending on the degree of comprehension required, as well as on the specific demands of the text (Bonk 2000, Stæhr 2009, van Zeeland and Schmitt 2013, Teng 2016). Though written language is hardly comparable to spoken language, these findings are in line with those for written language (Laufer 1989, Hu and Nation 2000, Laufer and Ravenhorst-Kalovski 2010, Schmitt *et al*. 2011). (See section 2 of this paper for a full discussion of these studies.)

5.4 Psycholinguistic validity of lexical counting unit

One final point which needs to be addressed is the psycholinguistic validity the lexical counting units used in this study. Using word families as the lexical counting unit to establish how much vocabulary is needed to understand English is based on the assumption that if a person knows one of the members of the word family then they will also be able to understand other inflected and transparently derived forms of the word. Research indicates that this may be the case, at least, for adults and L2 learners with high general English language proficiency (see Gardner 2007 for a review), particularly for the receptive skill of listening comprehension, which is the focus of the present study. However, it would not hold true for the productive skills of speaking and writing, for which research (Schmitt 1997, Schmitt and Zimmerman 2002) indicates that the flemma or word type may be more suitable lexical counting units.

Though this study makes no claims for productive English language usage, figures are also provided for the lexical coverage of VOICE not only of word

families, but also of flemmas and word types. The aim of providing also this information is to be more transparent about what any particular number of word families might translate into in terms of the number of word types or flemmas. The analysis of lexical coverage in VOICE revealed that to reach 95% lexical coverage in VOICE, 1,204 word families, 1,633 flemmas or 2,598 individual word types are needed (see Figure 1). Instead, to reach 98% coverage, 2,242 word families, 3,259 flemmas and 5,278 word types are required.

One limitation of the lexical counting units used in this study concerns multi-word units which form semantically inseparable units, but which can be difficult to identify and process electronically. These include open compounds (*air conditioning*), phrasal verbs (*put up with*), idioms (*rock the boat*), fixed expressions (*good afternoon*), and prefabs (*the point is*). Resource and time constraints for the current project meant that is was not feasible to count such items as single, holistic lexical units. Instead, they were counted separately as individual words and placed under their respective word families.

It is not clear what impact this operationalisation might have on the psycholinguistic validity of the construct of the lexical counting used in study. On the one hand, there is substantial evidence that, at least, native speakers of English (for a review see Schmid 2017: 7) store and access these multi-word lexical items (semi-)holistically, without the need for online composition. However, for non-native speakers the evidence is mixed, with only proficient users showing signs of some or partial holistic representation and retrieval (for a review, see Conklin and Schmitt 2012). Thus, it is possible that the operationalisation of the lexical counting unit used in this study leads to more psycholinguistically valid estimation of the lexical learning burden for L2 learners of English.

Another psycholinguistic factor considered in the operationalisation of the lexical counting unit used in this study concerns words with the same form but multiple meanings: for example, *bank* meaning a financial institute or the side of a river. The intended meaning of such homonyms is generally made clear by their context and L2 learners of English are likely to perceive them as separate words in their respective contexts. However, in a machine-based count such homographic words are indistinguishable. Thus, if adjustments are not made such lexical counts can lead to an underestimate of the learning burden for L2 English language learners.

In this study, all identified homonyms were placed under separate word families (see Table 3). For example, the adjective *appropriate*, adverb *appropriately* and noun *appropriateness* forms found in VOICE were placed under one word family and the verb form *appropriated* formed a distinct, single constituent word family. It is, nonetheless, possible that not all homonyms present in VOICE were identified, in which case, the resulting lexical coverage figures may be marginally smaller than they would otherwise be. It is, nevertheless, unlikely that any such oversights would have any significant impact of the results of this study.

5.5 Function words *versus* content words in VOICE

In the profiling analysis of VOICE against the BNC lists, I also compared the proportion of *function* to *content* words. This analysis of the data revealed that a very high proportion (51%) of the tokens occurring in VOICE at the first 1,000-word level of the BNC word family lists were function words. This may seem surprising to readers not familiar with corpus linguistics, but it is actually a common finding, and one which upon reflection can be easily understood: function words are the structural components of language and are needed to form any utterance. Instead, content words convey meaning, and as such they are context dependent. In other words, content words are as diverse as the number of meaningful messages that humans wish to encode in words and convey to one another.

6 Conclusion

The findings of this study suggest that half as much vocabulary is needed to understand spoken English in international contexts where it is used as a lingua franca compared to what is needed in intranational contexts where it is used between native speakers of English. If 98% lexical coverage is assumed to be the required amount of vocabulary knowledge, then 3,000–4,000 word families (plus proper nouns) would suffice in ELF settings, compared to 6,000–7,000 word families (plus proper nouns) found by Nation (2006) for native-speaker discourse. This is good news for L2 learners of English who need to understand English in such ELF settings, as it represents a significant saving in vocabulary size targets for such learners.

This study is based on the largest freely available dataset of general, spoken ELF discourse in Europe, the VOICE corpus. Future studies should aim to supplement the findings of the present study with the lexical coverage figures also for ELF used in other parts of the world, such as Asia, or Latin America. For example, a sister corpus, the Asian Corpus of English (ACE), also exists, so a comparison study could be carried out to supplement the findings of this study. Moreover, it would be useful to investigate how lexical coverage figures might vary depending on other variables too, such as specialised genre, discourse domain (e.g. professional, educational or leisure), speech-event type (e.g. conversation, panel discussion, business meeting, etc.). Another interesting question is whether the presence of native speakers in the interaction affects the range of vocabulary used, and if so, in what way. Finally, even though it was found in this study that the sample size used did not significantly affect the results, it may still be useful to validate the finding of this study against a considerably larger corpus of general ELF data than is currently available.

References

ACE. 2014. The Asian Corpus of English. Director: Andy Kirkpatrick; Researchers: Wang Lixun, John Patkin, Sophiann Subhan, last accessed 4 February 2018, from http://corpus.ied.edu.hk/ace/

Adolphs, S. and Schmitt, N., 2003. Lexical coverage of spoken discourse. *Applied Linguistics*, 24(4), 425–438.

Albrechtsen, D., Haastrup, K., and Henriksen, B., 2008. *Vocabulary and writing in a first and second language: Process and development.* Basingstoke: Palgrave Macmillan.

Alderson, J. C., 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment.* London; New York, NY: Continuum.

Bauer, L. and Nation, P., 1993. Word families. *International Journal of Lexicography*, 6(4), 253–279.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E., 1999. *Longman grammar of spoken and written English*. London: Longman.

Bonk, W. J. 2000. Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14(1), 14–31.

Britannica Academic. *English language 2018.* Last accessed on 4 February 2018, from http://academic.eb.com.libproxy.helsinki.fi/levels/collegiate/article/English-language/109779.

Carlisle, J. F., 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal*, 12(3–4), 169–90.

Carver, R. P., 1994. Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior*, 26(4), 413–437.

Clahsen, H., Felser, C., Neubauer, K., Sato, M. and Silva, R., 2010. Morphological structure in native and non-native language processing. *Language Learning*, 60(1), 21–43.

Cobb, T., 2000. The compleat lexical tutor [Computer software].

Cook, V., 1999. Going beyond the native speaker in language teaching. *TESOL Quarterly* 33(2), 185–209.

Crystal, D., 2006. Chapter 9: English worldwide. In: D. Denison and R. Hogg (eds.) 2008. *A history of the English language*, Cambridge: Cambridge University Press, 420–439.

European Commission. 2016. *Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy*. Luxembourg: Eurostat.

Firth, A., 1996. The discursive accomplishment of normality: On "lingua franca" English and conversation analysis. *Journal of Pragmatics*, 26(2), 237–259.

Gardner, D., 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2), 241–265.

Gagné, C., Psycholinguistic approaches to morphology. *Oxford Research Encyclopedia of Linguistics*. Last accessed on 4 February 2018, from http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-258.

Gilner, L., 2016. Identification of a dominant vocabulary in ELF interactions. *Journal of English as a Lingua Franca,* 5(1), 27–51.

Google Ngram Viewer. 2012. Last accessed on 4 February 2018, from https://books.google.com/ngrams.

Goulden, R., Nation, P., and Read, J., 1990. How large can a receptive vocabulary be? *Applied Linguistics*, 11(4), 341–363.

Hardie, A., Baker, P. and McEnery, T., 2006. *Glossary of corpus linguistics.* Edinburgh: Edinburgh University Press.

Henriksen, B., Albrechtsen, D. and Haastrup, K. 2004. The relationship between vocabulary size and reading comprehension in the L2. *Angles on the English-speaking World*, 4(1), 129–140.

Holmes, J., Vine, B. and Johnson, G. 1998. *The Wellington Corpus of Spoken New Zealand English: A users' guide.* Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Hu, M. H. and Nation, P., 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–30.

Hunston, S., 2012. Pattern grammar. *The Encyclopaedia of Applied Linguistics*. Blackwell Publishing Ltd.

Jenkins, J., 2006. Points of view and blindspots: ELF and SLA. *International Journal of Applied Linguistics*, 16(2), 137–62.

Jenkins, J., Cogo, A. and Dewey, M., 2011. Review of developments in research into English as a lingua franca. *Language Teaching*, 44(3), 281–315.

Kachru, B., 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In: R. Quirk and H. G. Widdowson (eds.) *English in the world: Teaching and learning the language and literatures*. Cambridge: Cambridge University Press, 11–30.

Laufer, B., 1989. What percentage of text-lexis is essential for comprehension. In: C. Laurén, and M. Nordman (eds.), *Special language: from humans thinking to thinking* machines. Clevedon [England]: Multilingual Matters Ltd, 316–323.

Laufer, B., 1992. How much lexis is necessary for reading comprehension? In: P.J.L. Arnaud and H. Bejoint (eds.), *Vocabulary and Applied Linguistics.* London: Macmillan, 126–132.

Laufer, B. and Goldstein, Z., 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning,* 54(3) 399–436.

Laufer, B., Elder, C, Hill, K., and Congdon, P., 2004. Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21, 202–226.

Laufer, B. and Ravenhorst-Kalovski, G., 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language,* 22(1), 15–30.

Laufer, B. and Yano, Y., 2001. Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, 13(2), 549–566.

Mauranen, A., 2011. Learners and users–Who do we want corpus data from. In: F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds.) 2011. *A Taste for corpora: In honour of Sylviane Granger.* John Benjamins Publishing, 45: 155–171.

Mauranen, A., 2017. A glimpse of ELF. In: M. Filppula, J. Klemola, A. Mauranen and S. Vetchinnikova (eds.), *Changing English: Global and Local Perspectives.* de Gruyter Mouton, 92: 223–253.

Milton, J. and Alexiou, T., 2009. Vocabulary size and the common European framework of reference for languages. In: B. Richards, M. Daller, D.D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds.) *Vocabulary studies in first and second language acquisition: The interface between theory and applications.* Houndmills Basingstoke: Palgrave Macmillan, 194–211.

Milton, J., Wade, J. and Hopkins, N., 2010. Aural word recognition and oral competence in English as a foreign language. In: R. C. Beltrán, C. Abello-Contesse, and M. del Mar Torreblanca-López (eds.) *Insights into non-native vocabulary teaching and learning.* Multilingual Matters, 52: 83–98.

Nagy, W. E., Diakidoy, I. N. and Anderson. R. C., 1993. The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Reading Behavior,* 25(2), 155–80.

Nation, I. S. P., 1993. Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing*, 10(1), 27–40.

Nation, I. S. P., 2001. *Learning vocabulary in another language*. Ernst Klett Sprachen.

Nation, I. S. P., 2004. A study of the most frequent word families in the British National Corpus. In: P. Bogaards, and B. Laufer. (eds.) *Vocabulary in a second language: Selection, acquisition, and testing.* John Benjamins Publishing, 10: 3–13.

Nation, I. S. P., 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82.

Nation, I. S. P., 2016. *Making and using word lists for language learning and testing.* Amsterdam: John Benjamins Publishing Company.

Nation, I. S. P., and Meara, P., 2000. Vocabulary. In: N. Schmitt (ed.) *An introduction to applied linguistics.* Cambridge: Cambridge University Press, 34–52.

Nation, I. S. P. and Waring, R., 1997. Vocabulary size, text coverage and word lists. In: N. Schmitt, and M. McCarthy (eds.) *Vocabulary: description, acquisition and pedagogy.* Cambridge: Cambridge University Press, 2035: 6–19.

Nation, I. S. P. and Webb, S. A., 2011. *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Nelson, C., 1995. Intelligibility and world Englishes in the classroom. *World Englishes*, 14: 273–279.

OED Online. 2018. Oxford University Press. Last accessed on 4 February 2018, from http://www.oed.com.libproxy.helsinki.fi/.

Pitzl, M., Breiteneder, A. and Klimpfinger, T., 2008. A world of words: processes of lexical innovation in VOICE. *Vienna English Working Papers* 17(2), 21–46.

Qian, D., 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308.

Qian, D., 2002, Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52: 513–536.

Ravin, Y. and Leacock, C., 2000. Polysemy: An overview. In: Y. Raven and C. Leacock (eds.) *Polysemy: Theoretical and computational approaches.* Oxford: Oxford University Press, 1–29.

Read, J., 1988. Measuring the vocabulary knowledge of second language leamers. *RELC Journal*, 19(2), 12–25.

Read, J., 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10: 355–371.

Read, J., 1998. Validating a test to measure depth of vocabulary knowledge. In: A. Kunnan (ed.), *Validation in language assessment.* Mahwah, NJ: Erlbaum, 41–60.

Read, J., 2004. Research in teaching vocabulary. *Annual Review of Applied Linguistics*, 24: 146–61.

Schmid, H. J., (ed.) 2017. *Entrenchment, memory and automaticity: The psychology of linguistic knowledge and language learning*. Berlin, Boston: De Gruyter Mouton.

Schmitt, N. 1997. Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition,* 19(1), 17–36.

Schmitt, N., 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., 2008. Review article: Instructed second language vocabulary learning. *Language Teaching Research,* 12(3), 329–363.

Schmitt, N., 2010. *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N., Cobb, T., Horst, M. and Schmitt, D., 2017. How much vocabulary is needed to use English? Replication of van Zeeland and Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, 50(2), 212–226.

Schmitt, N., Jiang, X. and Grabe, W., 2011. The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43.

Schmitt, N. and Marsden, R., 2006. *Why is English like that? Historical answers to hard ELT questions.* University of Michigan Press.

Schmitt, N. and Meara, P., 1997. Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, 19: 17–36.

Schmitt, N., Schmitt, D., and Clapham, C., 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18: 55–89.

Schmitt, N. and Zimmerman, C. B., 2002. Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145–171.

Schonell, F. J., Meddleton, I. G and Shaw B. A., 1956. *A study of the oral vocabulary of adults.* Brisbane: University of Queensland Press.

Seidlhofer, B., 2001. Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11(2), 133–158.

Seidlhofer, B., 2004. Research perspectives on teaching English as a Lingua Franca. *Annual Review of Applied Linguistics*, 24: 209–39.

Simon, E. and Taverniers, M., 2011. Advanced EFL learners' beliefs about language learning and teaching: A comparison between grammar, pronunciation, and vocabulary. *English Studies*, 92(8), 896-922.

Sinclair, J., 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J., 2005. Corpus and text—basic principles. In: M. Wynne (ed) *Developing linguistic corpora*: *A guide to good practice*. Oxford: Oxbow Books, 92: 1–16.

Smith, L. E., 1978. Some distinctive features of EIIL vs ESOL in English language education. *Cultural Learning Institute Report*, 5(3), 5–11.

Stæhr, L. S., 2008. Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal,* 36(2), 139–152.

Stæhr, L. S., 2009. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577–607.

Teng, F. 2016. An in-depth investigation into the relationship between vocabulary knowledge and academic listening comprehension. *TESL-EJ*, 20(2), 1–17.

Tyler, A. and Nagy, W., 1989. The acquisition of English derivational morphology. *Journal of Memory and Language*, 28: 649–67.

Tyler, A. and Nagy, W., 1990. Use of derivational morphology during reading. *Cognition,* 36: 17–34.

van Zeeland, H. and Schmitt, N., 2012. Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics,* 34(4), 457–479.

VOICE. 2013. *The Vienna-Oxford International Corpus of English* (version POS XML 2.0). Director: Barbara Seidlhofer; Researchers: Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn.

VOICE. 2013. Corpus Information, Vienna-Oxford International Corpus of English, last accessed on 4 February 2018, from https://www.univie.ac.at/voice/page/corpus_information.

VOICE. 2013. Corpus Description, Vienna-Oxford International Corpus of English, last accessed 4 February 2018, from https://www.univie.ac.at/voice/page/corpus_description.

Wang, Y., 2017. Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, 65: 139–150.

Widdowson, H. G., 1994. The ownership of English. *TESOL Quarterly*, 28(2), 377–389.

Widdowson, H. G. 2013., ELF and EFL: What's the difference? Comments on Michael Swan. *Journal of English as a Lingua Franca,* 2(1), 187–193.

World Bank. 2016. *World Development Indicators 2016.* Washington, DC, last accessed on 4 February 2018, from https://openknowledge.worldbank.org/handle/10986/23969.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A. and Healy, N. A. 1995. Growth of a functionally important lexicon. *Journal of Reading Behavior,* 27(2), 201–212.